A Comprehensive Survey on Computer Vision based Approaches for Automatic Identification of Products in Retail Store

Bikash Santra*, Dipti Prasad Mukherjee

Electronics and Communication Sciences Unit
Indian Statistical Institute
Kolkata, India

Abstract

The ability to recognize a product on the shelf of a retail store is an ordinary human skill. The same recognition problem presents an exceptional challenge for machine vision systems. Automatic detection of products on the shelf of a retail store provides enhanced value-added consumer experience and commercial benefits to retailers. Compared to machine vision based object recognition system, automatic detection of retail products in a store setting has lesser number of successful attempts. In this paper, we present a survey of machine vision based retail product recognition system and define a new taxonomy for this field. We also describe the intrinsic challenges associated with the problem. In this comprehensive survey of published papers, we analyze features used in state-of-the-art attempts. The performances of these approaches are compared. The details of publicly available datasets are presented. The paper concludes pointing to possible directions of research in related fields.

Keywords: Survey, product detection, product recognition, planogram compliance, multiple object detection, out-of-stock detection

1. Introduction

For long computer vision practitioners are attempting to build machine vision system to detect merchandise stacked in the racks of supermarket. By detection (or identification), we refer to recognition and precise localization of products visible in the racks of a supermarket. It is assumed that ideal marketing image of the individual product is available to the vision system. The objective of such a vision system is (1) to generate an inventory of products available in the store at any point of time from the images of racks stacked with products (referred as out-of-stock detection problem), (2) to correlate and validate the plan of product display (often referred as *planogram*) with the actual display of merchandise (referred as planogram compliance problem), and finally (3) to provide a value-added experience to users (referred as shopping assistance problem). In this paper we survey the progress of these product detection systems targeted for displayed merchandise in a supermarket.

The block diagram of the machine vision system under discussion is shown in Fig. 1. In rest of the paper, we interchangeably use rack image as shelf image and product image as product template. A set of typical product images from publicly available GroZi-120 dataset [1] is shown in Fig. 2(a). Example rack images where the product images of Fig. 2(a) are to be detected is shown in Fig. 2(b).

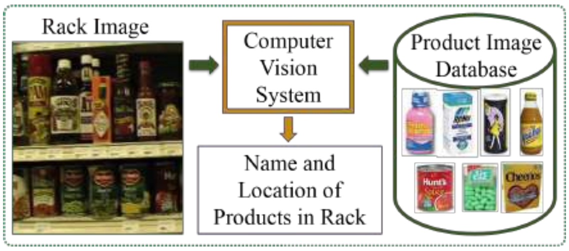
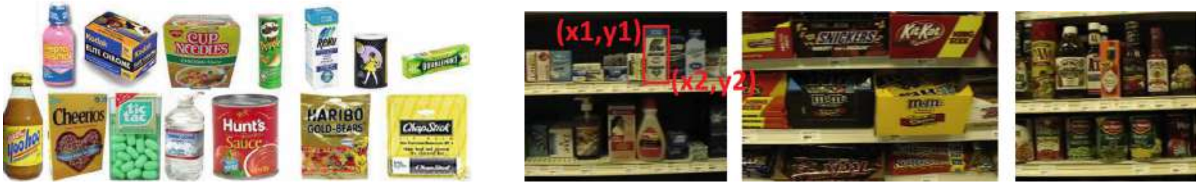


Figure 1: A typical computer vision system for detection of products in supermarkets

A few attempts have been made to solve the abovementioned problem using RFID, sensors, or barcodes [2, 3, 4]. There are ubiquitous sensor based system (like AmazonGo [5]) to monitor recognition and selection of products by a consumer. Most sensor based systems require fabrication at the manufacturer's end resulting in cost escalation of the product. Moreover, to assess the out-of-stock problem, a retailer needs to wait till the product leaves the store. Consequently, planogram compliance problem cannot be addressed with such sensors. Individual product based sensor has the problem of assessing the status of multiple products at one go. Devices for ubiquitous system have scalability issue and require significant investment. In contrast, computer vision based methods use hand phone camera or rack mounted camera to collect data. Overall,

*Corresponding author
Email addresses: `bikash.santra@isical.ac.in` (Bikash Santra), `dipti@isical.ac.in` (Dipti Prasad Mukherjee)



(a)

(b)

Figure 2: GroZi-120 dataset [1]: (a) sample product images typically used for marketing, (b) sample rack images where products are to be recognized and localized. $(x1, y1)$ and $(x2, y2)$ are the spatial co-ordinates of upper-left and bottom-right corners of a detected bounding box respectively.

computer vision based approaches provide an inexpensive feasible alternative compared to sensor based approaches.

In this paper, we present a comprehensive survey of the methods and results published over last 20 years in context of detection of products in retail stores. In a survey on object detection algorithms, Jafri *et al.* [6] present few methods of retail product recognition for the visually impaired persons. In a recently published conference article [7], authors present a brief survey on product recognition in shelf images. Their brief survey includes use of remote sensing technology for recognizing products in shelves. However, the survey in [7] is neither comprehensive nor explores the challenges specific to the problem. On the contrary, our comprehensive survey presents challenges, approaches, applications and new frontiers of the retail product detection problem. Furthermore, one of the important goals of this comprehensive survey is to present a new taxonomy of computer vision based state-of-the-art methods in detecting products on the shelves of a supermarket.

The rest of the paper is organized as follows. Section 2 discusses the challenges and benefits for automatic detection of products from images of racks of retail stores. Feature analysis is surveyed in Section 3. Section 4 presents an organized survey of methods. Sections 5 to 9 successively describe each group of works of the proposed taxonomy of our survey. Section 10 contains the description of publicly available datasets and comparison of performances of different methods based on published results. Finally, Section 11 concludes the paper pointing to future directions of research.

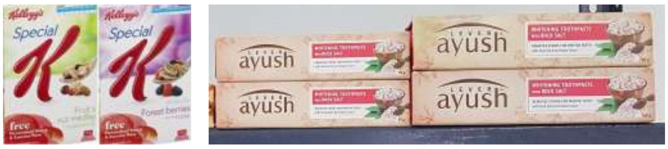
2. Challenges and Benefits

Table 1 summarizes the possible challenges of the product detection system.

Table 1: Challenges in automatic recognition of retail products

Category	Sub-category
Retail Store Environment	○ Complexity of scene ○ Data distribution ○ Variability of products ○ Fine-grained classification
Digital Imaging	○ Blurring ○ Uneven lighting conditions ○ Unusual viewing angle ○ Specularity

The racks are typically cluttered and often not organized in a regular fashion. Ideal marketing images of different products available to the vision system are often taken using different cameras resulting in different distributions of image intensities. Also, due to different imaging parameters, length of the product package (in some unit of length, say, cm) is mapped to different pixel resolutions for product and rack images. Examples of differences between product templates and rack images are evident in Fig. 2(a) and 2(b). Product packages come in different shapes and sizes. There are minor promotional variations in product packaging and a product detection system must differentiate such minor variations. This identification of minor signature variation in shape or color for a wide variety of products demand fine-grained classification. Fig. 3 demonstrates a few examples of visually similar products having minute changes in color, text or size.

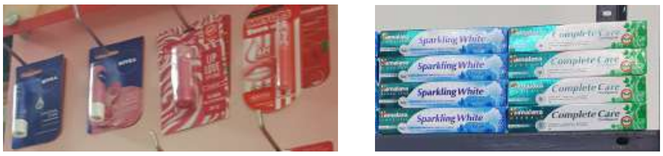


(a)

(b)

Figure 3: Fine-grained variations (a) color and text, and (b) size

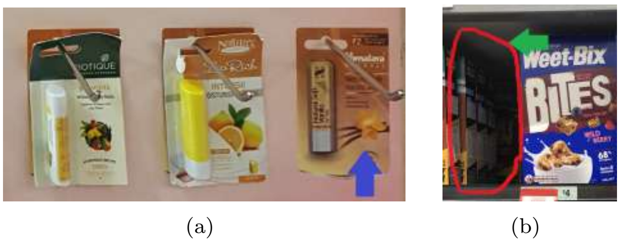
The rack images are captured using handheld devices. This often results in image blur due to camera shake and jitter (see the middle rack image in Fig 2(b)). The image of rack gets distorted due to oblique viewing (non-fronto-



(a)

(b)

Figure 4: (a) Distorted rack image due to oblique viewing and specular reflection due to glossy product packages and (b) rack image with vertically stacked products



(a) (b)

Figure 5: (a) The promotional sticker (marked by blue arrow) looks like a product and (b) shadow of a product (marked by the red contour and green arrow) in the gap

parallel position of the camera with respect to the rack) and uneven illumination (see Fig. 4(a)). The captured rack image often has specular reflection due to glossy product packages (see Fig. 4(a)). Object to image distortions magnify for stacked products (see Fig. 4(b) where parallel lines appear to generate a vanishing point due to oblique viewing) due to top and bottom boundaries of stacked product. The challenge often extends to a scenario where a gap in the rack (absence of product on the rack) is often get confused with the presence of a sticker in the peg board (see Fig. 5(a)) or with the presence of a product due to shadow and uneven illumination (see Fig. 5(b) marked using the red contour).

These reasons altogether pose significant challenge on top of typical object detection system studied in computer vision. The retail product detection problem bundles up various modalities of object detection problems like multiple object detection [8, 9, 10], detection of the multiple instances of the same object [11, 12], multiple object localization [13, 14], multi-view object detection [15], and fine-grained classification [16, 17, 18, 19].

The benefits of a vision based product detection system are summarized below.

1. *Enhanced Consumer Experience*: Note that there are around 30 million people in the world who are suffering from blindness [20]. Even for a normal buyer real time information of availability of a particular product at a given location of the store reduces the shopping time to a great extent.
2. *Commercial Benefits*: An estimate by Metzger *et al.* shows that out-of-stocks in supermarkets generally remain within a range of 5 to 10% [21]. In [22], Gruen *et al.* conduct a research on the impacts of out-of-stocks in retail stores worldwide. They find the following statistics due to out-of-stock situation: 31% shoppers move to another stores, 22% shoppers purchase another brand of the products, and 11% customers do not buy at all. The strategy for arrangement (planogram) of products in one or consecutive racks increases sales. Planogram establishes a close relation between shoppers, retailers, distributors, and manufacturers. It is observed that 100% optimized planogram compliance can increase sales up to 7 to 8% [23]. Hence, out-of-stock detection and checking of planogram compliance contribute to profit in retail businesses [24].

In the next section, we analyze features used in the attempts to recognize retail products.

3. Features for Detection of Retail Products

The feature descriptors for the problem under consideration are broadly classified as key point based, gradient based, pattern based, color based and deep learning based features. The related works under these classifications are tabulated in Table 2. Next, we present a brief discussion on each of the groups.

3.1. Key Point based Features

The key point based features are the most used for recognition of retail products. The retail merchandise are packaged in colorful and catchy outfits. As a result, the image of product package generates a number of key points suitable for image matching. The key points in most cases are detected using SIFT [25, 26] and SURF [39, 40]. The methods in this category that deserve special attention are [51, 52]. These approaches propose new variants of SURF namely AB-SURF [51] and NSURF [52] in detecting products. Overall Table 2 shows the importance of key point based features. Local image characteristics in and around key points are captured using a histogram [25, 26, 39, 40]. Stability of these histograms as features is one of the reasons for popularity of key point based features for detecting retail products.

Table 2: Feature descriptors and corresponding approaches where these features are used.

Categories	Feature Descriptors	Approaches
Key point based Features	SIFT [25, 26]	[1, 27, 28, 29, 30, 31], [32, 33, 34, 35, 36]
	Dense SIFT [37]	[38]
	SURF [39, 40]	[41, 42, 43, 44, 45], [46, 47, 48, 49, 50]
	AB SURF [51]	[51]
	Neo SURF [52]	[52]
	BRIGHT [53]	[54]
Gradient based Features	Morphological Gradient [55]	[56]
	HOG [57]	[58, 59]
	Sobel Operator [60]	[44]
	Canny Edge Detector [61]	[59]
Pattern based Features	Haar-like Features [62]	[1]
	Recurring Patterns [63]	[64, 65, 66]
Color based Features	Color Histogram [67]	[1, 41, 42, 31, 44]
	Saliency [68]	[51, 58, 49]
	Color Constancy Model [69]	[70, 71, 72, 73], [74, 75, 76]
Deep Learning based Features	CaffeNet [77]	[78, 79]
	AlexNet [80]	[50, 66]
	Inception-V3 [81]	[82]
	VGG-f [83]	[84]
	CNN [85, 86]	[49]

3.2. Gradient based Features

Gradient based features (e.g., HOG or Sobel operator) are used for template based matching of product images extracted from images of racks. The geometric shapes like corners or edges embedded in product and rack images are also utilized for template matching. Similarly, as in case of key point based features, gradient based local image characteristics are also captured using a histogram for detecting retail products [58, 59]. However, the performance of straightforward implementation of HOG like features are not encouraging as discussed in 5.1.

3.3. Pattern based Features

In identifying retail products, the most common pattern based features are Haar or Haar-like features [62] and recurring patterns [63]. In this category, the recurring patterns play a vital role in detecting products as in [64, 65, 66]. In many real-life situation, similar yet non-identical objects often appear in a group like cars on the street, faces in a crowd and in context of this paper, products on a supermarket rack. The authors of [63] state that *much of our understanding of the world is based on the perception and recognition of shared or repeated structures.* In order to capture such repeated structures or recurrence nature, each product in a supermarket rack, act as a unit in a recurring pattern. Fig 6 demonstrates two example images of rack where the circles indicate recurring patterns. Recently, the authors of [66] utilize the concept of recurring patterns in their proposed solution for the problem.

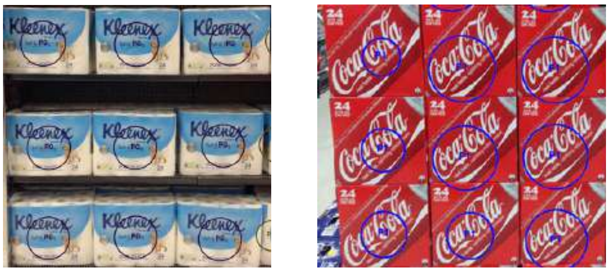

(a) (b)

Figure 6: Example images indicating recurring patterns by circles [64]: the images are taken from [64].

3.4. Color based Features

In detecting products, the color histogram [67] and classical saliency features [87, 68, 88] of products can be considered as color based features. However, saliency and color histogram are sensitive to illumination changes common to a retail store. In order to tackle such illumination effects in color images, the authors of [70]-[76] present various color based features using color constancy models for recognition of objects. List of color based features for product identification are given in Table 2.

3.5. Deep Learning based Features

In detecting retail products, all previously discussed four categories of features are hand crafted. In contrast, deep learning based features are derived from CNN pipeline [85, 86]. For retail product detection, either the outputs of an intermediate layer [50, 78, 66] of a network are used as features or the network as a whole is utilized for both feature extraction and classification [49, 84, 78, 79, 82]. Table 2 compiles deep learning related references. Next we present the taxonomy of the state-of-the-art methods of recognition system of retail products.

4. A Taxonomy for Detecting Retail Products

The first serious attempt [70] of recognition of retail products in isolation (i. e., identification of individual product image cropped from the rack image) was in 1999. Naturally, localization issue is not addressed in this work. It took almost another eight years to take a more involved approach for recognition and localization of multiple retail products. In 2007, Merler *et al.* [1] introduce the retail product detection problem along with a dataset containing rack and product images. Since then there are slightly more than 35 research publications directly related to retail product detection system. In Table 3, we propose a taxonomy for automatic detection of retail products.

From the pattern of development over the last decade, we find two major sequential steps as noted in Table 3. In the first layer of taxonomy, a probable region (containing a product) on the rack is identified based on an objectness (or productness) measure. We group the methods in the first layer in five different approaches: block, geometric transformation, saliency, detector, and user-in-the-loop based methods. Moreover, block based methods are classified into sliding window and grid based methods.

In the second layer of taxonomy, each method is partitioned into two groups namely unsupervised and supervised approaches of object detection. While using the terms supervised and unsupervised approaches, we have relied on the classical definitions used in the machine learning literature [91]. The unsupervised methods mainly include template based matching. The supervised methods refer to building a model using a training set. The trained model is used to test a new set of data unseen to the model.

The Table 3 also presents different areas of applications and corresponding categories of the problem. The areas of applications are (AI) Shopping assistive system (AII) Out-of-stock detection (AIII) Planogram compliance. The categories of the detection problem addressed in these papers are:

(DI) Detection of single product: This relates to accurate identification and localization of only one product at a time in a rack image.

(DII) Detection of multiple products: This relates to accurate identification and localization of all the products in a rack image in one go.

Table 3: Taxonomy of computer vision based approaches for product detection in retail stores (*: these methods crop product images from rack image either manually or using planogram information): for details, refer to text in Section 4.

			Unsupervised Methods	**Supervised Methods**	**Area of Application**	**Category of Problem**
Automatic Product Detection in Retail Stores	**Block based Methods**	**Sliding Window based Methods**	[1]		AI	DII
			[58]		AII	DII
			[44]		AIII	DII
			[52]		AIII	DII
		Grid based Methods	[28]		AI	DIV
			[29]		-	DIV
			[41]		AI	DI
			[54]		AII	DII
				[30]	AI	DII
				[89]	AI	DII
	Geometric Transformation based Methods		[1]		AI	DII
			[27]		-	DII
			[32]		AIII	DII
			[43]		AII	DII
			[45]		AI	DI
			[46]		AI	DI
			[47]		-	DII
			[34]		AIII	DII
			[35]		-	DII
			[48]		AI	DI
			[36]		AIII	DII
				[38]	AIII	DII
	Saliency based Methods		[42]		AI	DII
			[51]		AIII	DI
			[56]		AIII	DI
			[50]		AI	DII
				[49]	AI, AIII	DII
				[50]	AI	DII
				[66]	-	DII
	Detector based Methods			[1]	AI	DII
				[31]	AIII	DII
				[59]	AIII	DII
				[84]	-	DII
	User-in-the-loop Methods		[64]*		AIII	DIII
			[65]*		AIII	DIII
			[70]		-	DIII
			[71]		-	DIII
			[72]		-	DIII
			[73]		-	DIII
			[74]		-	DIII
			[75]		-	DIII
			[76]		-	DIII
				[90]*	-	DIII
				[33]*	-	DIII
				[78]*	AI	DIII
				[79]*	-	DIII
				[82]	AIII	DIII

(DIII) Recognition of products: This relates to recognition or classification of isolated products where localization is not important.

(DIV) Retrieval of rack images: Given a pool of rack images, the goal is to retrieve the rack images containing the query product.

A comparative study on performances of these approaches under different categories of the problem (DI, DII, DIII and DIV) is presented in Section 10. Note that out of all state-of-the-art methods for detecting retail products, only five methods [48, 36, 64, 65, 33] assume the presence of planogram in order to locate the products in a rack. In these methods, planogram informs the algorithm about the particular product expected in a given location of the rack. Naturally, test for absence or presence of the expected product at a given location reduces the challenge of discovering a product in absence of planogram information. Next we discuss and assess each group of the taxonomy in detail. We start with the first approach, block based methods.

5. Block based Methods

In block based methods, several overlapping or non-overlapping blocks are selected from the rack image as potential regions containing products. Consequently, local features (like SIFT [25, 26], and SURF [39, 40]) are computed from each such block and also from each of the product templates. For each block of rack image, the features are matched with those of product images. The product image with highest matching score is selected as the product for the block. The final detection result is generated after applications of various post processing techniques [50].

As mentioned earlier, the block based methods are classified into two categories: (a) Sliding window, and (b) Grid based methods. Next we discuss these methods in detail.

5.1. Sliding Window based Methods

A graphical illustration of sliding window based methods is presented in Fig. 7. In this case, all state-of-the-art methods are unsupervised recognition techniques as described next chronologically.

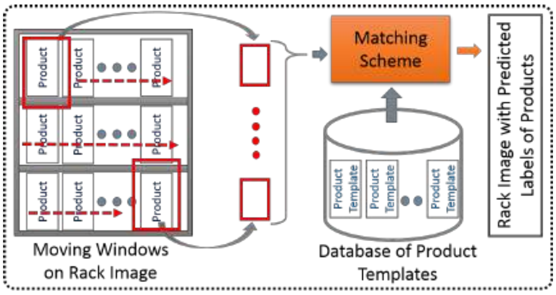

Product
Matching Scheme
Rack Image with Predicted Labels of Products
Product Template
Moving Windows on Rack Image
Database of Product Templates

Figure 7: Block diagram of a sliding window based method

Unsupervised Methods: In [1], Merler *et al.* provide three baseline approaches for their own dataset (made publicly available). Their primary goal is to address the recognition problem where the quality of product template (referred as *in vitro* image) largely differs from the rack image (referred as *in situ* image). They compare three kinds of localization approaches. One of them is the sliding window based approach. In this method, the probable product locations (i.e. regions of rack image) are localized by sliding windows of different scales. The concatenated histograms of *a* and *b* color planes of each region in *Lab* colorspace [92] is matched with that of product templates. The matching score of a probable region is determined by taking the smallest order statistic of intersections of histograms of the region and that of product templates.

For verification of planogram compliance, Marder *et al.* propose a solution in [58]. They detect products through two successive layers. In the first layer, they detect products in the rack image using three different methods: (i) point based vote map [93], (ii) sliding window based Histogram of Oriented Gradients (HOG) [57], and (iii) sliding window based Bag of Words (BOW) [94]. It is shown that vote map produces better result compared to other two methods. In the second layer, they refine the recognition performances by resolving visual ambiguity using saliency map. For the same planogram compliance problem, Saran *et al.* [44] provide a solution first finding the shelves (i.e. the horizontal bundle of lines in a rack image) utilizing Sobel operator followed by Hough transform [95, 96]. In the detected shelves, the products are localized and recognized through two consecutive steps: (i) sliding window based SURF feature correspondence, and (ii) false positive removal using color histogram matching between image patch in the rack and product images. Even though this method is evolved for verification of planogram compliance, the planogram information is nowhere utilized in the entire scheme.

Recently in [52], Ray *et al.* solve the planogram compliance problem using a two-layer approach. The first layer finds out exhaustive match of product images with the products in a shelf image by sliding a number of windows. The exhaustive list of sizes of sliding windows depend on the physical dimensions of products and shelf. Windows of the shelf image are matched with the product images integrating a variant of SURF and correlation scores. The second layer determines winner match out of all possible matches at a particular location using a graph theoretic approach.

5.2. Grid based Methods

The overview of grid based methods is graphically demonstrated in Fig. 8. As per the proposed taxonomy, state-of-the-art methods (see Table 3) under this category are described in two sub-categories: unsupervised and supervised methods as presented below chronologically.

Unsupervised Methods: Zhang *et al.* [28] introduce a baseline method for their dataset where detection of retail product is posed as an image retrieval problem. Their objective is to retrieve the rack images which contain a

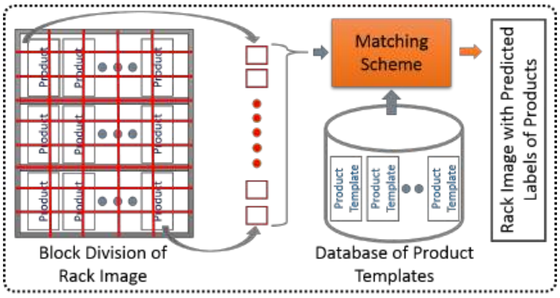

Product
Matching Scheme
Rack Image with Predicted Labels of Products
Product Template
Block Division of Rack Image
Database of Product Templates

Figure 8: Block diagram of a grid based method

given product template. They first apply a Harris-Affine interest region detector [97] to identify interest regions of rack images. From the interest regions, the SIFT features are extracted. Subsequently, a vocabulary is constructed with the help of SIFT features of rack images. In order to build the vocabulary, the visual words are determined by applying hierarchical k-means clustering [98] on SIFT features of all rack images. Next, any product or a region of rack image is represented using the histogram of these visual words. In retrieving rack images for a given product, all rack images are divided into a number of sub-images. Consequently histogram of visual words of each sub-image is matched with that of the given product image using four different similarity measures. The performances of four similarity measures are also compared in [28]. In 2009, Zhang *et al.* present an extension of this method in [29]. In their extension, they create the visual dictionary by considering different scales of their rack images. Further, an improved matching scheme is introduced to find a correct match between the block of rack and product images.

In [41], Bigham *et al.* propose a shopping assistive system (*VizWiz::LocateIt*) for visually impaired persons. In order to operate *VizWiz::LocateIt*, visually impaired persons are supposed to capture images of shelves when moving through an aisle and query for a specific product. The request is send to the remote server for addressing the query. In this system, the authors first divide the rack image into number of sub-images. Consequently, each sub-image is matched with product images. They present a comparison of two matching schemes: SURF [39, 40] and color histogram based matching. In another attempt, Higa *et al.* [54] utilize compact binary descriptor (BRIGHT) [53] for identification of key points and extraction of visual features from rack and product images. For each key point correspondence between rack image and product images, the product center is estimated in the rack image using the method as in [99, 100]. Subsequently, multiple identical products are identified by grid based voting of possible positions of product center and recursive geometric verification scheme.

Supervised Methods: In [30], George *et al.* present a multi-label image classification approach for localization and recognition of products. They first establish a locality-constraint linear coding (LLC) [101] model using dense SIFT features of product images. Consequently, they subdivide the rack image into several blocks. LLC features are then extracted from each block of the rack image and product images. A discriminative random forest [102] is trained with LLC features of product images. Using the trained model, a multi-class ranking of products is determined by classifying each block of the rack image. Furthermore, the authors perform a deformable spatial pyramid based fast dense pixel matching [103] and genetic algorithm based optimization scheme [104] for localization and recognition of products in the rack image. The authors of [30] make their dataset of products and rack images publicly accessible. In order to assist shoppers, George *et al.* [89] design a machine vision system which automatically finds out location of products in a rack image. The system has two separate modules. In the first module, to infer the information related to product, the brand name is predicted through two successive steps: (i) detection of text region using [105] and (ii) detection of texts in those regions using Optical Character Recognition (OCR) [106]. Remaining module takes care of fine-grained classification of products as follows: the visual features are extracted using discriminative patches as in [107]; the products are localized using spatial pyramid based image representation [94]; the products are recognized using SVM; finally the recognition performance is improved using active learning [108] through user feedback.

Pros and Cons: The primary advantages of block based methods are that the schemes are simple and easy to implement. The critical disadvantage is: how to choose the number and size of overlapping or non-overlapping blocks? In most cases, authors have chosen these parameters either experimentally or from prior knowledge. Thus, accurate localization of products can not be guaranteed in many cases. Moreover, the overlapping block based methods are computationally expensive.

The block based methods consider enormous number of sliding windows of different scales and sizes to locate the products in a rack. In other words, these methods exhaustively search for the products in the rack. As a result, these methods are robust against rotation and scaling of products in the rack.

On a different point, the slow execution of these methods is a major drawback in designing a real time system like shopping assistive application. To avoid exhaustive search for products in a rack, the geometric transformation based matching or graph theoretic approach looks like a promising direction of research. Next section presents the geometric transformation based methods.

6. Geometric Transformation based Methods

In retail store setting, images of racks captured by a handheld device undergo geometric transformation due to oblique view of camera with respect to the rack. As a result approaches in this group attempt to calculate features

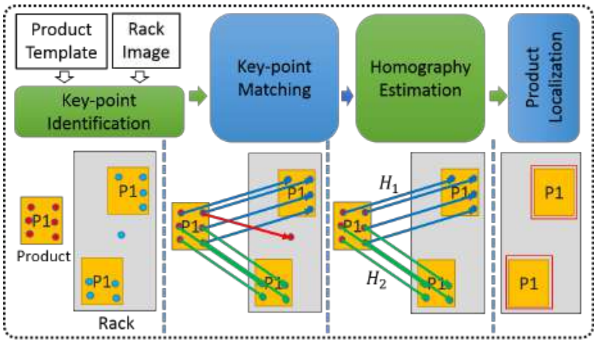


Figure 9: Block diagram of a geometric transformation based method: colored dots denote the key points, P1 represents a product and H_1, H_2 are the homographies between the product P1 & rack.

which are invariant to affine or projective object to image transformation. Most of the approaches in this group evaluate key point based local features (using SIFT, SURF etc.) for the rack and product images. The key point correspondences between rack image and product images are obtained using various techniques like clustering of key points or Hough voting. Finally, using these key point correspondences, products are recognized and localized in the rack image. In Fig. 9, we demonstrate a typical geometric transformation based method. As per the proposed taxonomy, state-of-the-art methods (see Table 3) under this category are described next in two sub-categories: unsupervised and supervised methods.

Unsupervised Methods: The block based matching approach in [1] described earlier in Section 5, utilizes geometric transformation. In this approach, the SIFT features are extracted from both rack and product images. Consequently, the correspondences of SIFT key points between rack and product images are determined. Using those matched key points, the homography matrices are calculated in order to locate the products in rack as described in [26]. In [27], Auclair *et al.* extract SIFT key points along with feature descriptors from product images and rack image. Consequently, they find correspondences of SIFT key points between product images and rack image. SIFT correspondences are determined by representing SIFT based features into two following data structures: k-d tree [109] and Locality Sensitive Hashing (LSH) [110]. In case of k-d tree based representation, Best-Bin-First algorithm [111] is applied to find the matches of key points. It is shown that LSH shows better performance compared to k-d tree. Finally, the products are detected in rack by calculating affine transformations of products using RANSAC [112] on the matched key points of rack and products. In [32], Bao *et al.* propose a Scale and Rotation Invariant Implicit Shape Model (SRIISM) in order to recognize and localize retail products. SRIISM is a modified version of Implicit Shape Model (ISM) [113]. In SRIISM, SIFT and BRIGHT [53] algorithms are applied to extract key points, scales and feature descriptors. At each key point of rack image, the features are matched with that of product images. Once the matching is complete, a probabilistic voting using Hough transform is carried out on matched pairs of key points (between rack image and product images) for the probable center, scale and orientation of the products. Consequently, the localization and pose of products are estimated by Best-Bin-First algorithm and maximum likelihood estimation respectively.

In [43], Kejriwal *et al.* aim to detect out of stock (OOS) situations in supermarkets. In their scheme, the OOS is detected by counting the products in a rack image. The products are recognized using k-d tree [109, 111] based representation of SURF feature descriptors of all product images. In the rack image, the products are counted in two ways: (i) computing the maximum repeatability of product image features, and (ii) localizing products using SURF correspondences along with RANSAC [112]. In contrast, Alhalabi *et al.* [45] develop a system to assist visually impaired shoppers. The proposed system asks for audio input from user for the brand name of a desired product. Consequently, the audio input is converted to text and a OCR technique is applied on product templates in order to select the product template for input product brand. Next, the system takes rack image as input for finding out the desired product brand. The product is recognized and localized by finding out SURF key point correspondences between rack image and product images followed by calculating the homography matrices of matched key points. In a similar context, the authors of [46] aim to develop a smart-glass based shopping assistive system for visually impaired persons. They match SURF key points of product images with that of rack image and calculate homography matrix to locate products in a rack image.

In [47], the authors determine locations of products in rack image by estimating poses of products. The pose is estimated with a Hough voting scheme. The Hough voting is conducted on matched SURF key points of rack image and product images. For recognition of products, a k-d tree [109] is constructed with SURF features of all product images. The products in the rack image are identified by applying Best-Bin-First search algorithm on previously constructed k-d tree. Furthermore, the fine-grained recognition is performed using pose-class histogram in Hough space. For detection of multiple instances of a product, Zhang *et al.* [34] present a dual-layer density estimation scheme. From all product images, SIFT feature descriptors are extracted to form a tree or hash representation. In their paper, the performance of different feature representation schemes (like k-d tree, hierarchical k-means tree [114, 115], vocabulary tree [116, 117], local sensitive hashing (LSH) [118, 119], Semi-supervised hashing (SSH) [120], and near-optimal hashing [121]) are compared. Among all those feature representation schemes, near-optimal hashing technique shows better accuracy and efficiency for their proposed model. For recognition of products, SIFT key points of rack image are matched with those of product images stored in near-optimal hash table. Consequently, the products are localized by applying a two-stage adaptive kernel density estimation method on matched key points as in [122]. A similar kind of approach is proposed in [35]. Zhang *et al.* [35] form an LSH representation of SIFT features of product images. For recognition of products, a nearest neighbor retrieval process is performed to find the correspondences between key points of rack image and those of product images stored in an LSH table. Subsequently, the products are localized by calculating a homography matrix of matched key points with the help of RANSAC algorithm.

In [48], Zientara *et al.* develop an integrated product detection system using android-powered smart glass[1] equipped with a camera and audio channel. The system is named as *Third Eye* which is developed to assist visually impaired shoppers in supermarkets. In this approach, the product is recognized by determining correspondences of SURF key points between rack image and product images. For single instance of products, the product is localized by calculating a homography matrix based on the correspondences of key points between rack and product images. For multiple instances of products, the planogram information is used to localize the products followed by the recognition of products using key point correspondences between rack and products. Recently in 2017, Tonioni *et al.* [36] attempt to solve the problem through two successive steps. First they find probable matches of product images (say observed output) in a rack image using SIFT and Hough transform. Later they apply sub-graph isomorphism between observed output and actual output (given in a planogram) in order to find out missing items and to remove false matches in observed output.

Supervised Methods: In order to assist visually impaired shoppers, Cleveland *et al.* [38] develop a navigating robot. They extract shelves from rack image using morphological operation followed by edge-linking algorithm. For localization of products, they utilize 3D point cloud of rack image derived from RGBD sensor or from the integration of laser and camera output [123]. The products are recognized through Näive Bayes nearest neighbor classifier [124] trained with dense SIFT [37] features of product images.

Pros and Cons: Geometric transformation based methods typically assume that the key points are identified correctly and key point correspondences are established accurately. Naturally, the performance of the schemes discussed above are dependent on assumptions related to key points.

If the products displayed on a rack are planar, homography estimation is not strictly necessary. Also, SIFT and SURF features are not sensitive to affine transformations between product and rack images. Unfortunately in retail store setting it is difficult to ensure the correct estimation of key points. Key points in a rack image are often missed due to poor illumination. On the other hand, more than desired number of key points are detected in a noisy rack image with cluttered background. This yields many incorrect geometric transformations between products and rack images.

However, for correct estimation of geometric transformations, scaling and rotation issues between product and rack images are automatically addressed. The entire approach is fast and suitable for real time implementation. Overall, geometric transformation based methods are promising and can be integrated with other approaches for a reliable result. Next we present saliency based methods.

7. Saliency based Methods

Saliency based methods localize products in a rack image by utilizing saliency maps [87, 49], gradient image [56], or by finding out potential regions [50] of rack image. Once the salient region of a rack image is determined, the local features of those interest regions are calculated and matched with that of product images. The block diagram of a typical saliency based method is presented in Fig. 10. Next unsupervised and supervised methods of saliency driven approaches are discussed chronologically.

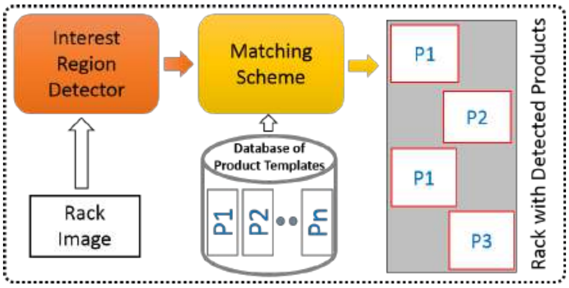


Figure 10: Block diagram of a saliency based method: P1, P2, $\cdots$, Pn are the products.

Unsupervised Methods: In [42], Winlock *et al.* try to solve the problem designing a shopping assistive system (named as *ShelfScanner*) for visually impaired shoppers. For any image, the feature vector is determined as the concatenated vector of all SURF descriptors and color histograms. Subsequently, dimension of the concatenated feature vector is reduced by applying Principal Component Analysis (PCA) [125]. Finally, a probabilistic framework is proposed to find out salient regions (i.e. locations of products) in the rack image. In a similar context, Thakoor *et al.* [51] aim to design a wearable aid for visually impaired shoppers. They introduce a variant of SURF, Attention Biased SURF (AB-SURF) features. An attention biased saliency map of the rack image is derived for each product image by the attention biasing algorithm [87]. Consequently, SURF features are extracted from salient regions and matched with SURF features of product images. In contrast, Frontoni *et al.* [56] build a smart camera for verification of planogram compliance. In their scheme, the morphological gradients [55] are calculated for the entire

[1] https://www.xda-developers.com/android-based-smart-glass-round-up-whats-new-at-ces-2016/ accessed as on 3rd Feb, 2019

rack image. The gradient image is then matched with product images using a template matching scheme as described in [126]. In [50], Franco *et al.* present a scheme using Harris corner detector [127] and 3D color histograms of rack image in YC_bC_r color space [128] to find out salient regions (i.e. possible locations of the products) in rack image. The visual features of each salient region of rack are compared with that of product images in order to recognize the products. Bag of Words (BoW) (i.e. a visual dictionary [129, 130]) based visual features is formed by applying k-means clustering [131] on SURF features of product images. The histogram of visual words from the visual dictionary uniquely represents any product image.

Supervised Methods: For visually impaired shoppers, Zientara *et al.* [49] present a product detection system using a smart glass similar as in [48]. In [49], the locations of products in a rack image are determined by deriving the saliency map of the rack image. The saliency map is computed using attention by information maximization (AIM) [88] and SURF key points of the rack image. Finally, the products are recognized using a CNN [85, 86], trained on their own dataset of product images. In [50], Franco *et al.* first transform the rack image into YC_bC_r color space [128]. The corners (using Harris corner detector [127]) and 3D color histograms of the transformed image are used in order to find out salient regions (i.e. possible locations of products) in a given rack image. CNN features from each probable location of rack image and product images are extracted using a pre-trained CNN model (AlexNet [80]). Subsequently, the Euclidean distances between the CNN features of each potential location and those of product images provide a matching score.

Recently in [66], the authors present a conditional random field (CRF) [132] based scheme. Similar products occupy neighboring locations in a rack. For example, bottles of soft drinks may be visualized as a linear chain. The local visual features are derived from convolutional layers of a CNN model. Given a sequence of products in a rack, the CNN features are extracted from the sequence and fed to a CRF model as a linear chain. Finally, forward-backward [133] and Viterbi [134] algorithms are applied on the CRF to find out labels of the given product sequence.

Pros and Cons: Saliency based methods are two-layered. In the first layer, the salient regions are identified in a rack image. The second layer matches the salient regions with products. In most cases, the first layer of these methods do not miss to identify regions containing products. But at the same time, the first layer tends to over-estimate the salient regions. As a result the saliency based localization methods usually fail when rack image contains partially-occluded products.

The second layer minimizes false detection due to cluttered background of the rack image. Like block and geometric transformation based methods, the salient region based methods also take care of rotations and scaling of products in rack. The implementation of second layer is relatively fast as the recognition is executed only for the salient regions. Shopping assistive system implemented with any of these methods can always be operated in real time.

Newer salinecy based deep learning tools like R-CNN [135], Fast R-CNN [136], Faster R-CNN [137], Mask R-CNN [138], and SSD [139] are yet to be explored for the problem under consideration. These methods require training data comprises of annotated rack images where each product is labeled with bounding boxes. However, in a retail store environment, capturing images of racks and annotating the same for building a significant training dataset, is a painstaking activity. In contrast, the template of the new packaging of a product is made publicly available before its actual arrival at the store. Therefore, template driven approaches are preferred for detecting products in supermarket.

Overall, saliency based methods require attention for detecting partially-occluded products, which is a normal situation in a retail store. Next section presents detector based methods.

8. Detector based Methods

For various real world objects like face [140] or pedestrian [141], there exists reliable dedicated detectors. Detector based methods (e.g. [1]) separately train a machine learning tool (like AdaBoost [142]) with certain domain-specific visual features (e.g. Haar-like features [62]) of product images to find out bounding boxes of products in the rack image. Once the bounding boxes are detected, the local visual features are extracted from the regions of rack image for matching with the product images. Fig. 11 demonstrates a graphical illustration of detector based methods. As per the proposed taxonomy, state-of-the-art methods (see Table 3) under this category are all supervised approaches, which are detailed next.

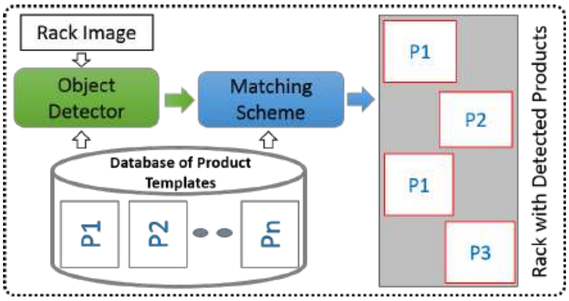


Figure 11: Block diagram of a detector based method: P1, P2, $\cdots$, Pn are the products.

Supervised Methods: The approach of [1] is discussed in Section 5. There are three methods in [1]. Out of these three, one is the detector based method. However, the authors of [1] experimentally show that geometric transformation based approach outperforms other two approaches namely sliding window and detector based methods. In [31], Varol *et al.* apply an object detector [140] to determine the bounding boxes of products in the rack image. After detecting the bounding boxes, the products are recognized by SVM [143] classifier which is trained with the color histograms and SIFT features of product images. A similar approach is presented in [59]. Rather than detecting bounding boxes of products from the entire rack image, the authors of [59] first detect shelves in rack image using Canny edge detector and Hough transform. Consequently, the product locations (i.e. bounding boxes) are detected from each shelf by the AdaBoost based modified Viola-Jones algorithm [140, 144]. In the detected locations, the products are categorized by SVM classifier trained with HOG and color features.

Recently in late 2017, Karlinsky *et al.* [84] present a product identification method consisting of three successive phases. In the first phase, they propose a probabilistic inference model based on the SIFT features of a number of patches of product images (i.e. training images) in order to find out initial set of detections. In the second phase, the initial set of detections are refined for fine-grained classification of products. In order to do that, the detected boxes from first phase are again classified through a re-trained CNN model (VGG-f network [83]). Finally, in the last phase, [84] integrates first and second phases with the KLT tracker [145] in order to track the detected boxes in a video.

Pros and Cons: Like saliency based methods, these are also two-layered approaches. First layer detects the bounding boxes using a object detector while the second layer takes care of the classification of products. The object detector trained with partially-occluded objects can determine the bounding boxes for partially-occluded products in a rack. The object detector of the first layer could also identify the product class for a bounding box. Utilizing this class information, the second layer may even perform finer classification of products (e.g. challenges shown in Fig. 3). However, this is not yet explored for methods in this category. The object detector trained with rotated and scaled objects could be a promising application. These approaches are suitable for real time applications.

In retail store setting, the intensity distributions of train and test images are not necessarily identical (as discussed in Section 2). As a result, the detector based approaches, especially those using learning schemes like AdaBoost, may fail to identify the bounding boxes. However, the statistical learning based object detector as in [84] does not suffer seriously from such problems. Next we present user-in-the-loop based methods.

9. User-in-the-loop Methods

User-in-the-loop methods do not automatically localize products in the rack. In the rack image, products are cropped out either manually or by utilizing product's information provided in a planogram. Subsequently, local features of cropped products are matched with that of product images. Thus, user-in-the-loop methods do not address primary challenges of localization of products in the rack. In the following paragraphs, the unsupervised and supervised approaches are chronologically briefed.

Unsupervised Methods: Approaches in [70]-[76] have provided color constancy model to detect individual retail products. Though not explicitly mentioned, examples show that the product images are cropped from rack images manually. Their proposed color constancy and shape context model able to recognize individual products inspite of uneven illumination and changes in color profile.

In order to verify planogram compliance, Liu *et al.* [64, 65] consider the problem as recognition of recurring patterns. A brief discussion on recurring patterns [63] is presented during feature analysis in Section 3.3 of this paper. The rack image is first partitioned into distinct regions utilizing prior information about products given in a planogram. In each region, the recurring patterns are recognized for detecting similar yet non-identical products. Consequently an estimated layout of shelves is generated displaying the detected products. For planogram compliance, the estimated layout is matched with the expected layout using spectral graph matching as in [146].

Supervised Methods: In retail stores, similar products are usually placed adjacent to each other. Thus the resultant context information is important. The placement of products forms a spatially continuous structure in terms of brand and size. Advani *et al.* [90] propose a Visual Co-occurrence Network (ViCoNet) learning model which integrates the context information. In their scheme, first the products are manually cropped from the rack image. Consequently, the cropped products are recognized by the ViCoNet model. Similar as [90], Baz *et al.* [33] present a context-aware scheme in order to perform fine-grained classification of retail products. First, the planogram information is used to crop products from the rack image. Subsequently, products are recognized through two steps: (a) context-free classification with SIFT features, Bag of Words features and SVM classifier, and (b) context-aware classification using two probabilistic graphical models: hidden markov model [147] and conditional random field model [148]. The methods in [78, 79, 82] are evolved using deep learning algorithms. The authors of [79] manually crop the products from rack and classify cropped images using CNN model. The work in [82] presents a product classification scheme without considering the localization of products. For classification of retail products, the authors of [78] determine features from the last fully connected layer of CNN.

Pros and Cons: These methods do not localize products in a rack image. Only classification or recognition performances (not detection) are judged using these user-in-the-loop approaches. As a result, these methods always show better detection performances than other related methods. In a realistic store level scenario with difficult challenges like identification of multiple shelves in a rack or identification of rack start and rack end points in a rack image, user-in-the-loop looks like a promising approach. This approach has the potential to detect a novel product (not already available with product dataset) or to identify a gap (missing product) in the rack space. Both these applications of novel product identification and gap identification have major commercial impacts for retailers.

In the next section, we present a comparative study on the performances of state-of-the-art methods.

10. Comparisons of Retail Product Detection Methods

As mentioned earlier, there exists more than 35 published papers for detection of products in retail stores. We present the published results in order to compare the performances of the existing approaches. However, to compare the results, there are two major constraints: (a) differences in evaluation protocols, and (b) differences in set of product templates and set of rack images. The limited availability of benchmark public datasets is yet another bottleneck. We begin the comparison by presenting the details of public datasets.

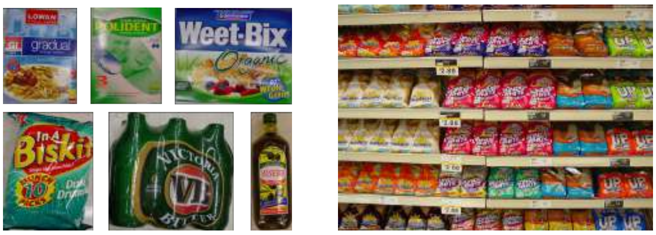
10.1. Publicly Available Datasets

Table 4 lists the publicly available datasets by their year of publications, number of product categories, number of product images and number of rack images. The datasets are briefly described in the following paragraphs.

Table 4: Publicly available datasets (*: 29 video footages, **: broad product categories)

Year	Dataset	# Product Categories	# Product Images	# Rack Images
2007	GroZi-120[2] [1]	120	676	29*
2007	WebMarket[3] [28]	100	300	3153
2014	Grocery Products [30]	27** / 3235	3235	680
2015	Grocery Dataset[4] [59]	10	3600	354
2016	Freiburg Groceries Dataset[5] [79]	25	4947	74

1) GroZi-120 [1]: GroZi-120 dataset is the first ever published benchmark dataset of grocery products. The important characteristic of this dataset is that the products and racks are imaged in completely different setup. Sample product and rack images from the dataset are shown in Fig. 2. The product images are collected from grocery web stores like Froogle[6]. The set of product images includes images with a variety of illuminations, sizes and poses as they are taken from different vendors or photo galleries. The rack images are captured in videos from retail stores at store-level variations in illumination, scale, reflectance, pose, color, and rotation. In addition, rack images have cluttered background. Videos of store shelves are recorded using a VGA resolution MiniDV camcorder at 30 fps. There are 29 videos with total duration of 30 minutes. The rack images are selected at every 5 frames of the videos. In GroZi-120 dataset, the product images are referred to as *in vitro* images while rack images are referred to as *in situ* images. The dataset is made of 120 categories of products. The number of instances for an *in vitro* image (i.e. product image) ranges from 2 to 14. Moreover, the dataset contains 14 to 814 *in situ* images corresponding to an *in vitro* image. For creating the ground truths, the rack image and location (i.e. bounding box) corresponding to each product image are manually annotated.

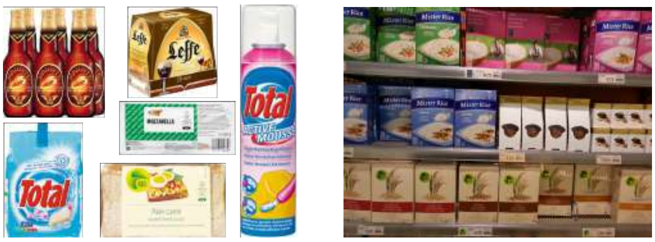


(a)

(b)

Figure 12: WebMarket dataset [28]: (a) product images (b) rack image.

2) WebMarket [28]: WebMarket dataset with images of size 2272 × 1704 or 2592 × 1944 are collected from 18 shelves each of length 30 meters in a retail store. The rack images are captured when the products are on the shelf. Each product is also captured off the shelf to use as product template. Thus, rack images differ from product images in scale, pose and illumination. Fig. 12 presents example of product and rack images. This dataset contains 3 instances for each of 100 product categories. The dataset also includes fine-grained product categories having minor variations in packages. The ground truth (as defined for the previous dataset) is generated by manually identifying location and product category of each product located in the rack images.



(a)

(b)

Figure 13: Grocery Products dataset [30]: (a) product images (b) rack image

[2] http://grozi.calit2.net/grozi.html accessed as on 3rd Feb, 2019

[3] http://yuhang.rsise.anu.edu.au/ accessed as on 3rd Feb, 2019

[4] https://github.com/gulvarol/grocerydataset accessed as on 3rd Feb, 2019

[5] https://github.com/PhilJd/freiburg_groceries_dataset accessed as on 3rd Feb, 2019

[6] http://www.froogle.com accessed as on 3rd Feb, 2019

3) Grocery Products [30]: The Grocery Products dataset is developed to address the fine-grained (i.e. similar yet non-identical) classification and localization of objects. The product images are collected from the web. The template images are captured in studio like environments. The rack images are captured using a mobile phone in natural retail store environment. Rack images are recorded from different viewing angles with various lighting conditions and magnification levels. In Fig. 13, we show a few samples of rack and product images from the dataset. The number of products in a rack image ranges from 6 to 30. The ground truth (as defined for the previous dataset) is manually created by labeling the categories and the locations (bounding boxes) of products in rack images. The dataset consists of 80 broad product categories. Out of those 80 categories, the ground truth is available only for 27 product categories under which 3235 fine-grained product templates are included.

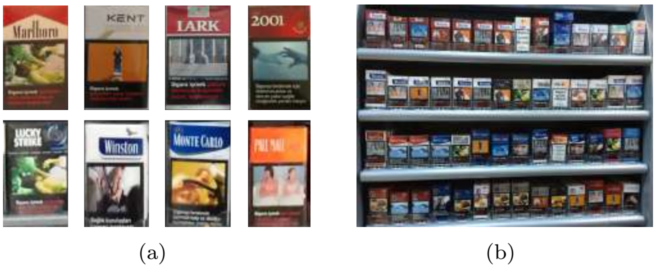

(a) (b)

Figure 14: Grocery Dataset [59]: (a) product images (b) rack image

4) Grocery Dataset [59]: In the Grocery Dataset, the product images (see Fig. 14(a)) are captured in a controlled environment with four types of cameras. The rack images (see Fig. 14(b)) are taken from 40 grocery stores with four types of cameras at various rack-to-camera distances. The number of products in a rack image ranges from 2 to 137. The dataset includes 200 instances (on an average) of 10 broad categories of products. In rack images, each product is annotated by covering the product with a bounding box using Google Image Clipper[7] tool (a framework for annotating ground truth of images). The ground truth is generated using the co-ordinates of the bounding boxes and corresponding product categories for each rack image.

5) Freiburg Groceries Dataset [79]: Freiburg Groceries Dataset consists of real world images of products and shelves. The product images are captured with four different cameras at some grocery stores, apartments, and offices in Freiburg, Germany. The dataset includes 97 to 370 instances of size 256 × 256 for each of 25 categories of products. In this dataset, the product images have cluttered background and variation in illumination. On the other hand, the rack images are recorded with a Kinect v2 camera[8]. Hence, the dataset comes with RGB image, depth image, and a point cloud of the scene corresponding to each rack image of size 1920 × 1080. Furthermore, the rack image displays multiple instances of stacked products with a cluttered background. Fig. 15 demonstrates some example of product and rack images from the dataset. Next we present the comparison of the published results of existing methods.

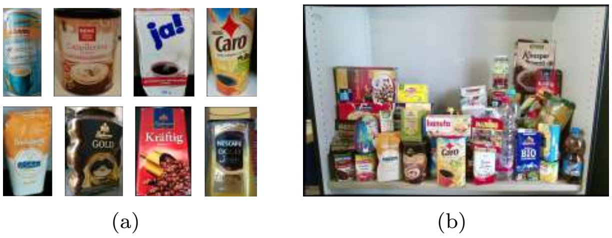

(a) (b)

Figure 15: Freiburg Groceries dataset [79]: (a) product images (b) rack image

10.2. Comparison using Published Results

In this survey, an extensive comparative study is conducted per category of the problem: (DI) detection of single product, (DII) detection of multiple products, (DIII) recognition of products, and (DIV) retrieval of rack images as listed in Table 3. Note that the results in rest of the tables and figures are reproduced from the respective publications. We also include the execution time (in seconds) of the product detection schemes and the corresponding computer configurations in rest of tables reproduced from the respective papers. In a few cases, we are unable to include results due to imprecise definitions of evaluation indicators [27, 38, 42, 46, 49]. Next we start our comparative study with detection of single product.

Table 5: Comparison of single product detection performances on private datasets: the values in bold indicate the best detection performances.

Publication	Detection Result		Data Specifications	
	Accuracy (%)	Recall (%)	#Products	#Racks
[41]	57.10	-	-	-
[45]$^{\Upsilon}$	78.00	-	775	1550
[51]$^{\Lambda}$	-	**94.12**	10	1037
[56]$^{\varrho}$	**85.00**	-	-	-

NOTE:

$^{\Upsilon}$Average results of two different categories of test images: rack with single instance and two instances of a product [45]

$^{\Lambda}$Average results of different product categories

$^{\varrho}$Results of their "MEDIUM" [56] case experiment

(DI) Detection of single product. We find that only four competing methods [41, 45, 51, 56] under this category for the comparisons. These methods are evaluated using accuracy and recall. The accuracy, recall and

[7] https://code.google.com/archive/p/imageclipper/ accessed as on 3rd Feb, 2019

[8] https://github.com/code-iai/iai_kinect2 accessed as on 3rd Feb, 2019

Table 6: Comparison of detection performances on benchmark datasets: the values in bold indicate the best detection performances.

Dataset	Publication	Detection Result			Data Specifications		Time Consumption	
		Accuracy (%)	Recall (%)	Precision (%)	#Products	#Racks	Test Time (sec/img)	Comp Specs
GroZi -120	[30]⊕	-	43.03	13.21	120	885	**01.95**	CPU: 2.4GHz RAM: 4GB
	[50] (BOW)	-	46.30	45.70	120	13120	-	-
	[50] (DNN)	-	**52.70**	45.20	120	13120	-	-
	[84]	-	-	**49.70**	120	4973	-	-
Grocery Products	[30]	-	68.50	30.70	3235	680	**01.95**	CPU: 2.4GHz RAM: 4GB
	[89]	-	61.90	-	3235	680	27.60	CPU: 3.4GHz RAM: 16GB
	[47]	**84.60**	-	-	3235	680	09.00	CPU: Core i7
	[36]	-	**90.20**	**90.40**	181	70	-	-
	[50] (BOW)	-	76.50	77.70	20	71	-	-
	[50] (DNN)	-	73.60	73.10	20	71	-	-
	[84]⊠	-	-	44.72	3235	680	-	-
	[52]	-	88.51	-	-	-	-	-
Grocery Dataset	[31]	-	89.00	**88.00**	10	229	-	-
	[59]	-	**94.00**	81.00	10	354	-	-
WebMarket	[52]	-	**90.8**	-	-	-	-	-

NOTE:
BOW: Bag of Words; DNN: Deep Neural Network
⊠In this paper, the "Grocery Products" dataset is referred as "GroZi-3.2K"

precision are calculated using True positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). As given in [45], the count of TP is increased when a product is present in the rack and the algorithm detects it. The count of FP is increased when the product is not present in the rack but the algorithm detects the product. The count of TN is increased when the product is not present in the rack and the algorithm also does not detect it. Finally, the count of FN is increased is the count of product present in the rack and the algorithm does not detect it. Consequently, the accuracy, recall and precision are defined as: Accuracy $= \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FN}+\text{FP}+\text{TN}}$, Recall $= \frac{\text{TP}}{\text{TP}+\text{FN}}$ and Precision $= \frac{\text{TP}}{\text{TP}+\text{FP}}$.

All of the methods [41, 45, 51, 56] are evaluated on the private (or in-house) datasets. The Table 5 presents the results of the methods taken from the respective papers. In terms of accuracy, saliency based method [56] outperforms other methods in detecting single product in a rack.

(DII) Detection of multiple products. In order to detect multiple products in one go, the localization of

Table 7: Comparison of detection performances on private datasets: the values in bold indicate the best detection performances.

Publication	Detection Result			Data Specifications		Time Consumption	
	Accuracy (%)	Recall (%)	Precision (%)	#Products	#Racks	Test Time (sec/img)	Comp Specs
[58]	87.40	-	-	223	240	-	-
[44]	92.00	-	-	25	-	¡01.00	CPU: Core i3 RAM: 2GB
[54]	-	**96.10**	**99.70**	65	24	**00.14**	CPU: 3.5GHz
[32]ϑ	-	90.00	95.00	221	87	¡00.40	CPU: 3.5GHz
[43] (Method 1)†	-	82.19	90.00	96	2681	-	-
[43] (Method 2)†	-	77.85	90.00	96	2681	-	-
[34]Ξ	-	95.40	95.30	70	-	02.90	CPU: 2.5GHz RAM: 8GB
[35]ϖ	**94.66**	-	-	-	-	00.43	CPU: 2.5GHz
[66]	-	87.00	91.00	972	24024	-	-
[84]	-	-	91.30	121	567	-	-
[52]	-	92.4	-	750	150*	-	-

NOTE:
ϑThe recall value is extracted from recall-precision graph
†Average results of 4 partitions of their dataset: D1, D2, D3, D4
ΞAverage results of partitions of their dataset: High, Medium, and Low texture level
ϖAverage results of partitions of the dataset: 5 categories of products
*Experiments are performed per shelf of the rack images (> 2000 shelf images)

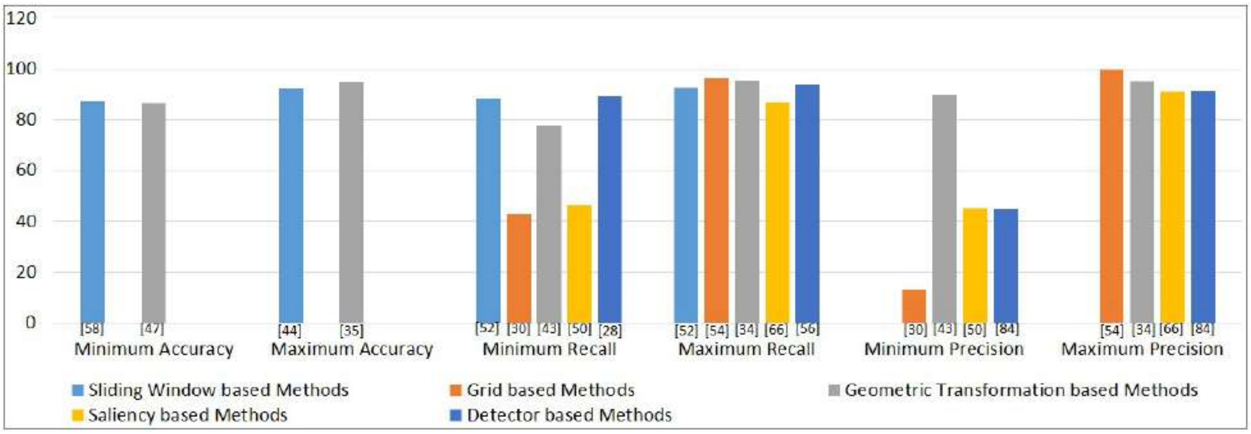


Figure 16: Performance analysis of proposed taxonomy: not all the references have reported accuracy, recall and precision measures.

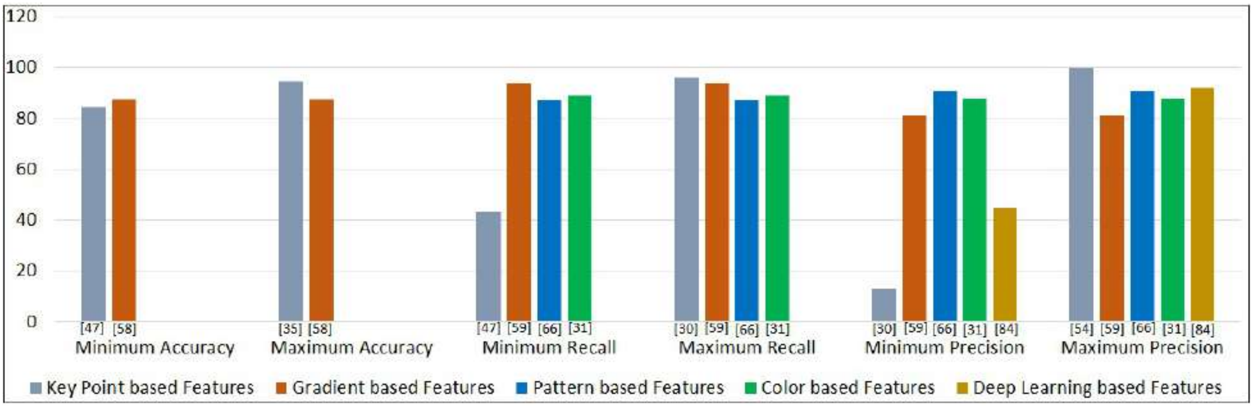


Figure 17: Performance analysis of features: not all the references have reported accuracy, recall and precision measures.

products is the primary step of the solution. Thus the taxonomy (as presented in 3) is proposed primarily based on the localization strategies of the solutions. The organization of our comparative study on this category of the problem is as follows: (a) first we discuss the evaluation indicators reported in the related papers. Subsequently, we present the dataset specific performances of state-of-the-art methods in Tables 6 and 7. (b) Next we present the best and worst performances of the proposed taxonomy over all datasets in Fig. 16. We also present the best and worst performances of the features over all datasets in Fig. 17. (c) The performances of supervised and unsupervised methods over all datasets are compared in Fig. 18. (d) Dataset-wise best methods are provided in Fig. 19. (e) Finally, we provide a few other performance indicators and the corresponding references (see Table 8). Highlights of each of the above points are briefly discussed in the following paragraphs.

Similar to single product detection system, the performance of multiple product detection system is measured using recall, precision and accuracy as described in [50]. The definitions of recall, precision and accuracy remain same as in case of single product detection system. However, the definitions of TP, FP, FN and TN vary as follows. If the center of a detected product box lies within the ground truth product box (or the intersection over union between the detected and ground truth product boxes ¿= 50%) and the label of the detected product is same as that of ground truth, then the detected product is considered as a TP. If the label of the detected product is different from that of the ground truth, the detected product is then considered as an FN. If the center of the detected product box is outside of the ground truth product box (or the intersection over union between the detected and ground truth product boxes ¡ 50%), then the detected product is considered as an FP. Otherwise the detected product is considered as a TN.

Table 6 lists the benchmark dataset specific results reproduced from respective papers while Table 7 presents the results on private datasets. In some state-of-the-art methods, the recall is calculated as detection accuracy over the entire test dataset. In that case, we report the recall rate as detection accuracy in Table 6 and Table 7.

From Fig. 16, it can be noticed that the geometric transformation based methods show consistent performance for all the indicators. Note that the Fig. 16 does not include user-in-the-loop methods as they only deal with recognition of products and not their localization.

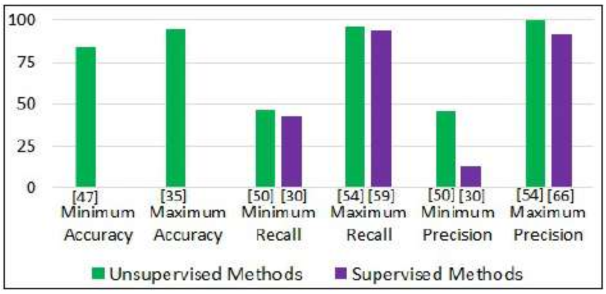


Figure 18: Performance analysis of unsupervised and supervised methods: performance is not reported using accuracy measure for supervised methods.

From Fig. 17, the superiority of key point based features is clearly established. However, note that deep learning based features are still not widely explored in the field of retail product detection. Fig. 18 shows that the unsupervised methods outperform supervised methods in case of all the indicators due to unavailability of large number of instances of products. The deep learning based supervised schemes [59, 66] show almost equal performances compared to unsupervised schemes.

In Fig. 19, we present the best performance on each dataset. It is to be noted that neither recall nor precision rate exceeds 60% on GroZi-120 dataset. Although the performance on private (refer ‘Others’ in Fig. 19) dataset is significant, that is not the case for challenging public datasets.

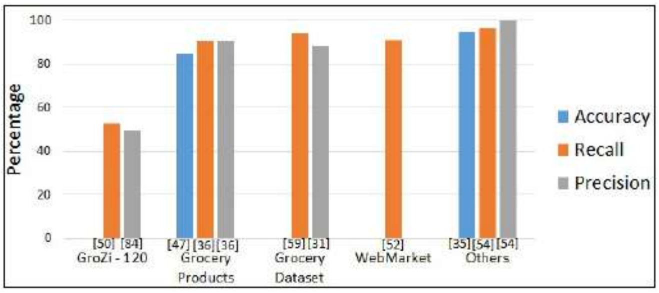


Figure 19: Best results on each dataset

As suggested in [50], the performance of a retail product detection system can also be measured by overall recall and overall precision rate. These measures are utilized for evaluating localization performances of a product detection system. The recall of a true product can be determined as the proportion of the ground truth area covered by the detected boxes. The precision of detected product can be calculated as the proportion of the detected product area that covers the ground truth product. Thus, for any given rack image I, the recall and precision can be defined as:

○ $\text{Recall}(I) = \frac{1}{|G^I|} \sum_{G_r^I} \frac{\mathtt{A}(\cup G_r^I \cap \cup D_r^I)}{\mathtt{A}(\cup G_r^I)}$

○ $\text{Precision}(I) = \frac{1}{|D^I|} \sum_{G_r^I} \frac{\mathtt{A}(\cup D_r^I \cap \cup G_r^I)}{\mathtt{A}(\cup D_r^I)}$

where $\mathtt{A}(\cdot)$ represents the area of a spatial region, G_r^I represents the ground truth object of r-th class in I, D_r^I represents the detected object of r-th class in I, $\cup G_r^I$ is the spatial union of ground truth objects of r-th class in I, $\cup D_r^I$ is the spatial union of detected objects of r-th class in I, $|G^I|$ is the number of all ground truth objects in I, and $|D^I|$ is the number of all detected objects in I. Consequently, the overall recall and overall precision (as suggested in [50]) can be determined as the weighted mean of recall and precision respectively over all images in the test dataset T.

○ $\text{Overall Recall} = \frac{1}{|G|} \sum_{I \in T} |G^I| \cdot \text{Recall}(I)$

○ $\text{Overall Precision} = \frac{1}{|D|} \sum_{I \in T} |D^I| \cdot \text{Precision}(I)$

where $|G| = \sum_{I \in T} |G^I|$ and $|D| = \sum_{I \in T} |D^I|$ are the total number of ground truth objects and the total number of detected objects respectively in T. These two indicators are described as continuous measures.

Table 8: Comparison of detection performances measured by continuous measure

Dataset	Publication	Method	Detection Result	
			Overall Recall (%)	Overall Precision (%)
GroZi-120	[1]†	CHM	15.00	17.00
		SIFT	72.00	18.00
		Adaboost	15.00	17.00
	[50]$^\otimes$	BOW	41.80	**39.20**
		DNN	**44.40**	37.50
Grocery Products	[50]$^{\text{II}}$	BOW	**65.40**	73.70
		DNN	54.70	**73.90**

NOTE:
CHM: Color Histogram Matching
BOW: Bag of Words
DNN: Deep Neural Network
†evaluated on the entire dataset
$^\otimes$evaluated on 13120 #racks for 120 #products
$^{\text{II}}$evaluated on 71 #racks for 20 #products

In context of retail product detection, retailers are more interested in knowing the percentage of correctly identified products in rack rather than the ratio of areas of ground truth and detected products in a rack. Perhaps due to this reason, only two approaches [1, 50] evaluate the performance using overall recall and overall precision. The first paper on retail product detection, [1] presents the performances of the proposed system using these measures on the GroZi-120 dataset. Later, Franco *et al.* [50] also measure the performances of the proposed system using these measures on GroZi-120 and Grocery Products datasets. In Table 8, we tabulate the reproduced results from the respective papers. In case of GroZi-120 dataset, Franco *et al.* [50] establish the superiority of their proposed schemes (Bag of Words (BOW) and Deep Neural Network (DNN) based schemes) over the methods of [1]. In Grocery Prod-

Table 9: Comparison of recognition performances on one benchmark and other private datsets: the values in bold indicate the best detection performances.

Dataset	Publication	Detection Result			Data Specifications	
		Accuracy (%)	Recall (%)	Precision (%)	#Products	#Cropped Products
Freiburg Groceries Dataset	[79]	**78.90**	-	-	25	-
Private Datasets	[90]	73.00	-	-	62	¿1000
	[33] (HMM)	78.02	-	-	794	108090
	[33] (CRF)	**79.86**	-	-	794	108090
	[78]	63.00	-	-	5	-
	[82]	-	89.30	91.90	8	-

NOTE:
HMM: Hidden Markov Model; CRF: Conditional Random Field

ucts dataset, [50] achieves highest recall of 65.4% using BOW based approach and highest precision of 73.9% using DNN based approach.

(DIII) Recognition of products. The methods under this category deal only with recognition of products. The products are cropped from rack (manually or using planogram information) and the cropped products are recognized by the respective method. These methods are evaluated using accuracy, recall and precision. The definitions of these performance indicators are same as in case of (DI) detection of single product. The recognition performances of the methods are tabulated in Table 9.

In context of planogram compliance, the performance of product detection system can be evaluated using overall accuracy of planogram compliance [64]. Let d be the number of ground truth labels in a dataset and T be the test dataset. For each product label, let N_{det} be the number of detected products in a rack image $I \in T$. Let N be the number of products specified in the planogram for the rack image I. The planogram compliance accuracy (PCAc) for l-th product label is defined as:

∘ $\text{PCAc}(I, l) = 1 - \frac{|N_{det} - N|}{N}$.

Subsequently, the overall planogram compliance accuracy, OPCAc over all product labels and over all rack images can be defined as:

$$\circ\ \text{OPCAc} = \frac{1}{d \cdot |T|} \sum_{I \in T} \sum_{l=1}^{d} \text{PCAc}(I, l),$$

where $|T|$ is the number of rack images in the test set T.

Table 10: Comparison of performances of planogram compliance

Publication	OPCAc (%)	Data Specifications	
		#Products	#Racks
[64]	**91.47**	39	-
[65]	90.53	39	-

In the literature, we find only two published research papers reporting the overall accuracy of planogram compliance. In Table 10, we present the results from [64] and [65].

(DIV) Retrieval of Rack Images. In context of image retrieval, the performance of any retail product detection system is measured using image retrieval accuracy as defined in [28]. In this case, the objective of retail product detection system is to retrieve the rack images which display a given product. As suggested in [28], the image retrieval accuracy can be defined as the percentage of products that have true matches with rack images. Formally, the image retrieval accuracy can be defined as:

∘ Image Retrieval Accuracy $= \frac{c}{n}$,

where c is the number of instances of the query images retrieved in top s racks and n is the number of instances of the query images in the ground truth.

Table 11: Comparison of rack retrieval performances measured by image retrieval accuracy (IRA) on the entire WebMarket dataset

Publication	# of Top Ranked Retrieved Images (s)	IRA (%)	Time Consumption	
			Test Time (sec/image)	Computer Specs
[28]	50	70.00	-	-
[29]	5	**75.00**	0.025	CPU: 2.80 GHz RAM: 4 GB

In terms of image retrieval accuracy (IRA), the performances of state-of-the-art methods are presented in Table 11. It can clearly be seen that [29] outperforms [28] by exactly 5%. Not only that, [29] establishes its superiority within only 5 top retrieved rack images. Notably, the authors of [28] consider top 50 retrieved rack images in order to measure the performance of their proposed scheme. Next we summarize our survey pointing to important future directions of research.

11. Summary and Concluding Remarks

This paper presents the first comprehensive survey on state-of-the-art methods of automatic identification of products on display in a supermarket using computer vision based techniques. The performances of object recognition systems using computer vision is being pushed constantly for over last 50 years now [149]. In contrast, the com-

Table 12: Key takeaways from the survey

Best methods in different contexts		
Contexts	**Methods**	**Remarks**
Shopping assistive system	Saliency based methods [51, 56]	This is a single product detection problem. In Table 5, it can be seen that saliency based methods outperform other methods. These methods are suitable for real-time implementation.
For retailers	Block based methods [44, 54]	Provides solution for planogram compliance by detecting multiple products in one go. Problem of accurate localization of products is addressed. These are computationally expensive methods. But retailers can wait for the result at the cost of better accuracy.
Egocentric camera based application	Geometric transformation based methods [34]	Egocentric camera usually captures near-field images of rack. So the affine transformation between product templates and near-field image of rack can be determined accurately. However, as of now there is no related application using egocentric cameras.
Robot or drone mounted camera	Detector based methods [84]	Requires video based solution. Detector based methods should take care of relative motion and detection of partially occluded objects. Drone based retail product identification is a possibility yet to be explored.
Shelf-mounted cameras	Block based methods [44, 53]/ Detector based methods [84]	In this pictures are relatively stable and viewing angle is fixed. Naturally, both block and detector based methods should equally perform well within a reasonable execution time.
For detection of single product	Saliency based methods [51, 56]	In Table 5, it is evident that saliency based methods perform better compared to other groups of methods.
For detection of multiple products	Saliency based methods [66]	From Table 6 and Table 7, it can be seen that saliency based methods is not the best in detecting multiple products. However, these methods could lead to an initial estimate of products which may be accurately detected using machine learning approaches.
For recognition of products	Deep convolutional neural network [59, 66]	CNN based approaches are well known for these applications. Scalability, data augmentation for single instance of product templates and minimal re-training are key issues that need attention.
For retrieval of rack images	Block based methods [29]	Table 11 shows block based method is a good choice. However, only two attempts have been made to solve this problem.

Issues addressed successfully	
Issues	**Remarks**
Detection of single product when rack images are captured from near fronto-parallel viewpoint	This issue has been solved to a great extent as evident in the results in Table 5 to Table 11.
Camera for capturing product template and rack are unknown	This challenge is addressed to some extent as evident in the results in Table 5 to Table 11.

Issues that need attention	
Issues	**Remarks**
Fine grained classification of products	This particular issue is addressed only in [30]. Since then, we do not see any progress in addressing this challenge in detecting products.
Identification of gaps between products	No attempts have been made yet to address this problem.
Identification of novel products	No attempts have been made yet to address this problem.
Estimation of scale between products and rack	Only one attempt [58] has been made to estimate the scale between products and a rack. This scale information is a fundamental prerequisite for product recognition.
Identification of vertically stacked products	No attempts have been made yet to address this problem.
Identification of products with uneven illuminations and specular effects	Few attempts have been made to solve these issues representing images in different color spaces like *Lab* [1] and *YCbCr* [50]. Still these are important challenges.
Detection of products in rack image captured using non-fronto parallel camera	There is no state-of-the-art approach that explicitly addresses this problem.

munity is not that active in the particular problem space discussed in this paper.

There are clearly two views to the problem under discussion. From the consumer's point of view, product recognition and localization are the key challenges. While product recognition was addressed by many, localization and assessment of localization accuracy need attention.

From the retailer's point of view, product recognition, localization and planogram compliance are important. At the same time, automatic identification of gap (empty rack space) is equally important and this particular problem needs serious research attention.

In Table 12, we assess key takeaways from the survey. The table answers the following questions. (a) Which methods are best for different application scenarios? (b) What are the key issues that have been solved? (c) What are the remaining issues that need more attention from computer vision practitioners?

Given these, the natural question is: what are the key system characteristics of an automatic computer vision application for identifying products displayed on image? The answer is the following.

11.1. Desirable Key System Characteristics

- **Real-time**: From the consumer perspective, the system must operate in real-time such that availability of products can be checked immediately. From the retailer's stand point, the system should operate close to real-time given that the store must respond to consumer need for replenishing the stock. From the perspective of system integrator, there is always a trade off between processing the image at the hand held device (where the rack image is captured) or at the background or cloud. However, system implementation is yet to receive serious attention from the researchers.
- **Accuracy**: The system should consistently operate at high level of accuracy for a wide range of products in order to establish its acceptability within the consumers and retailers. There should be minimum or no user interaction.
- **Robustness**: The key challenges come from mismatch in scale between a product template and the rack image, uneven illumination, variation in camera angle, and unstable image capturing due to hand held devices. From the machine learning stand point (especially considering deep learning architecture), the major bottleneck is the availability of single instance of the product image. Naturally, synthetic generation of training images using data augmentation technique requires special attention for machine learning based technique to improve its performance.

11.2. Future Directions

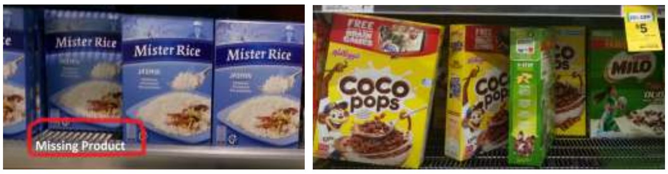
Looking at the challenges discussed so far, following few key directions of research has emerged.

(a) Successful generation of region proposals: Key point or saliency based region proposals that can successfully crop a potential region containing a product is the key to the success of a learning model. Proposals of multiple regions and discovery of their arrangements on the racks at one go using a graph theoretic or a constraint optimization approach should be important research challenges.

(b) Semantic Segmentation: Potential regions may be discovered from certain definition of objectness followed by assigning them a class label. Non-maximal suppression of objectness could identify semantic segments. Each pixel [150] of the rack may be labeled for detecting the products. Recent success of deep semantic segmentation models [151, 152, 153] should provide appropriate motivation.

(c) Scalability and adaptability: Product packages change quickly. Number of products increases rapidly. Frequent retraining of machine learning system is a serious bottleneck. These issues present interesting research problems. A related difficult research problem is novelty detection when a very similar but new product needs to be identified as novelty compared to the existing product templates.

(d) Logo and OCR: Both recognition of logo and optical character recognition are popular research topics and have seen limited success in many applications. There is no serious effort as of now to integrate logo and OCR technologies with retail product recognition system.



(a) (b)

Figure 20: Examples of (a) missing product (marked by red contour) and (b) tilted products on the shelf due to manual mishandling

(e) Utilization of depth information: Use of RGBD [154] based system needs to be explored especially to identify gaps and missing products (see Fig. 20(a)). Appearance of gaps is often influenced by uneven illumination. Therefore, integration of RGBD system with other appropriate sensor data may be explored.

(f) No dedicated algorithm is explored for stacked products, especially when the products are arranged vertically (see Fig. 4(b)). Similarly, recognition strategies need to be made robust to certain extent such that manual mishandling of products (see Fig. 20(b)) can be taken care of.

Overall, these challenges and approaches suggest that the retail product identification system will continue to be an exciting research and development field with sufficient room for improvement in the years to come.

References

[1] M. Merler, C. Galleguillos, S. Belongie, Recognizing groceries in situ using in vitro training data, in: Computer Vision and

Pattern Recognition, 2007. CVPR'07. IEEE Conference on, IEEE, 2007, pp. 1–8.

[2] D. López-de Ipiña, T. Lorido, U. López, Indoor navigation and product recognition for blind people assisted shopping, in: International Workshop on Ambient Assisted Living, Springer, 2011, pp. 33–40.

[3] V. Kulyukin, A. Kutiyanawala, Accessible shopping systems for blind and visually impaired individuals: Design requirements and the state of the art, The Open Rehabilitation Journal 3 (2010) 158–168.

[4] J. Nicholson, V. Kulyukin, D. Coster, Shoptalk: independent blind shopping through verbal route directions and barcode scans, The Open Rehabilitation Journal 2 (1) (2009) 11–23.

[5] T. Bishop, How amazon goworks: the technology behind the online retailers groundbreaking new grocery store, GeekWire. Extraído de https://www. geekwire. com/2016/amazon-go-works-technology-behind-online-retailersgroundbreaking-new-grocery-store.

[6] R. Jafri, S. A. Ali, H. R. Arabnia, S. Fatima, Computer vision-based object recognition for the visually impaired in an indoors environment: a survey, The Visual Computer 30 (11) (2014) 1197–1222.

[7] C. G. Melek, E. B. Sonmez, S. Albayrak, A survey of product recognition in shelf images, in: Computer Science and Engineering (UBMK), 2017 International Conference on, IEEE, 2017, pp. 145–150.

[8] B.-N. Vo, B.-T. Vo, N.-T. Pham, D. Suter, Joint detection and estimation of multiple objects from image observations, IEEE Transactions on Signal Processing 58 (10) (2010) 5129–5141.

[9] M. Villamizar, A. Garrell, A. Sanfeliu, F. Moreno-Noguer, Interactive multiple object learning with scanty human supervision, Computer Vision and Image Understanding 149 (2016) 51–64.

[10] K. Oh, M. Lee, G. Kim, S. Kim, Detection of multiple salient objects through the integration of estimated foreground clues, Image and Vision Computing 54 (2016) 31–44.

[11] Z. Haladová, E. Šikudová, Multiple instances detection in rgbd images, in: International Conference on Computer Vision and Graphics, Springer, 2014, pp. 246–253.

[12] G. Aragon-Camarasa, J. P. Siebert, Unsupervised clustering in hough space for recognition of multiple instances of the same object in a cluttered scene, Pattern Recognition Letters 31 (11) (2010) 1274–1284.

[13] H. He, S. Chen, Imorl: Incremental multiple-object recognition and localization, IEEE Transactions on Neural Networks 19 (10) (2008) 1727–1738.

[14] G. L. Foresti, C. Regazzoni, A change-detection method for multiple object localization in real scenes, in: Industrial Electronics, Control and Instrumentation, 1994. IECON'94., 20th International Conference on, Vol. 2, IEEE, 1994, pp. 984–987.

[15] A. Torralba, K. P. Murphy, W. T. Freeman, Sharing visual features for multiclass and multiview object detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (5) (2007) 854–869.

[16] Z. Ge, A. Bewley, C. McCool, P. Corke, B. Upcroft, C. Sanderson, Fine-grained classification via mixture of deep convolutional neural networks, in: Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on, IEEE, 2016, pp. 1–6.

[17] H. Yao, D. Zhang, J. Li, J. Zhou, S. Zhang, Y. Zhang, Dsp: Discriminative spatial part modeling for fine-grained visual categorization, Image and Vision Computing 63 (2017) 24–37.

[18] T. Sun, L. Sun, D.-Y. Yeung, Fine-grained categorization via cnn-based automatic extraction and integration of object-level and part-level features, Image and Vision Computing 64 (2017) 47–66.

[19] D. Huang, R. Zhang, Y. Yin, Y. Wang, Y. Wang, Local feature approach to dorsal hand vein recognition by centroid-based circular key-point grid and fine-grained matching, Image and Vision Computing 58 (2017) 266–277.

[20] WHO, World health organization fact sheet, N° 282 (2014), accessed on 24-Jun-2017.
URL `http://www.who.int/mediacentre/factsheets/fs282/en/`

[21] C. P. Metzger, High fidelity shelf stock monitoring, Ph.D. thesis, ETH Zurich, Zurich, Switzerland (2008).

[22] T. W. Gruen, D. Corsten, S. Bharadwaj, Retail out of stocks: A worldwide examination of causes, rates, and consumer responses, Grocery Manufacturers of America, Washington, DC.

[23] M. Shapiro, Executing the best planogram, in: Professional Candy Buyer, Norwalk, CT, USA, 2009.

[24] M. O. M. Medina, Z. Fan, T. Ranatunga, D. T. Barry, U. Sinha, S. Kaza, V. Krishna, Customer service robot and related systems and methods, uS Patent App. 14/921,899 (Oct. 23 2015).

[25] D. G. Lowe, Object recognition from local scale-invariant features, in: Computer vision, 1999. The proceedings of the seventh IEEE international conference on, Vol. 2, Ieee, 1999, pp. 1150–1157.

[26] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International journal of computer vision 60 (2) (2004) 91–110.

[27] A. Auclair, L. D. Cohen, N. Vincent, How to use sift vectors to analyze an image with database templates, in: International Workshop on Adaptive Multimedia Retrieval, Springer, 2007, pp. 224–236.

[28] Y. Zhang, L. Wang, R. Hartley, H. Li, Where's the weet-bix?, in: Asian Conference on Computer Vision, Springer, 2007, pp. 800–810.

[29] Y. Zhang, L. Wang, R. Hartley, H. Li, Handling significant scale difference for object retrieval in a supermarket, in: Digital Image Computing: Techniques and Applications, 2009. DICTA'09., IEEE, 2009, pp. 468–475.

[30] M. George, C. Floerkemeier, Recognizing products: A per-exemplar multi-label image classification approach, in: European Conference on Computer Vision, Springer, 2014, pp. 440–455.

[31] G. Varol, R. S. Kuzu, Y. S. Akgiil, Product placement detection based on image processing, in: Signal Processing and Communications Applications Conference (SIU), 2014 22nd, IEEE, 2014, pp. 1031–1034.

[32] R. Bao, K. Higa, K. Iwamoto, Local feature based multiple object instance identification using scale and rotation invariant implicit shape model, in: Asian Conference on Computer Vision, Springer, 2014, pp. 600–614.

[33] I. Baz, E. Yoruk, M. Cetin, Context-aware hybrid classification system for fine-grained retail product recognition, in: Image, Video, and Multidimensional Signal Processing Workshop (IVMSP), 2016 IEEE 12th, IEEE, 2016, pp. 1–5.

[34] Q. Zhang, D. Qu, F. Xu, K. Jia, X. Sun, Dual-layer density estimation for multiple object instance detection, Journal of Sensors 2016.

[35] Q. Zhang, D. Qu, F. Xu, K. Jia, N. Jiang, F. Zou, An improved method for object instance detection based on object center estimation and convex quadrilateral verification, in: Information Technology, Networking, Electronic and Automation Control Conference, IEEE, IEEE, 2016, pp. 174–177.

[36] A. Tonioni, L. Di Stefano, Product recognition in store shelves as a sub-graph isomorphism problem, in: International Conference on Image Analysis and Processing, Springer, 2017, pp. 682–693.

[37] A. Bosch, A. Zisserman, X. Munoz, Image classification using random forests and ferns, in: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, IEEE, 2007, pp. 1–8.

[38] J. Cleveland, D. Thakur, P. Dames, C. Phillips, T. Kientz, K. Daniilidis, J. Bergstrom, V. Kumar, Automated system for semantic object labeling with soft-object recognition and dynamic programming segmentation, IEEE Transactions on Automation Science and Engineering.

[39] H. Bay, T. Tuytelaars, L. Van Gool, Surf: Speeded up robust features, Computer vision–ECCV 2006 (2006) 404–417.

[40] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (surf), Computer vision and image understanding 110 (3) (2008) 346–359.

[41] J. P. Bigham, C. Jayant, A. Miller, B. White, T. Yeh, Vizwiz:: Locateit-enabling blind people to locate objects in their environment, in: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, IEEE, 2010, pp. 65–72.

[42] T. Winlock, E. Christiansen, S. Belongie, Toward real-time grocery detection for the visually impaired, in: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, IEEE, 2010, pp. 49–56.

[43] N. Kejriwal, S. Garg, S. Kumar, Product counting using images with application to robot-based retail stock assessment, in: Technologies for Practical Robot Applications (TePRA), 2015 IEEE International Conference on, IEEE, 2015, pp. 1–6.

[44] A. Saran, E. Hassan, A. K. Maurya, Robust visual analysis for planogram compliance problem, in: Machine Vision Applications (MVA), 2015 14th IAPR International Conference on, IEEE, 2015, pp. 576–579.

[45] W. Alhalabi, D. Attas, Toward device assisted identification of grocery store sections and items for the visually impaired, in: Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV), The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2016, p. 49.

[46] R. Brenner, J. Priyadarshi, L. Itti, Perfect accuracy with human-in-the-loop object detection, in: European Conference on Computer Vision, Springer, 2016, pp. 360–374.

[47] E. Yörük, K. T. Öner, C. B. Akgül, An efficient hough transform for multi-instance object recognition and pose estimation, in: Pattern Recognition (ICPR), 2016 23rd International Conference on, IEEE, 2016, pp. 1352–1357.

[48] P. A. Zientara, S. Lee, G. H. Smith, R. Brenner, L. Itti, M. B. Rosson, J. M. Carroll, K. M. Irick, V. Narayanan, Third eye: A shopping assistant for the visually impaired, Computer 50 (2) (2017) 16–24.

[49] P. Zientara, S. Advani, N. Shukla, I. Okafor, K. Irick, J. Sampson, S. Datta, V. Narayanan, A multitask grocery assistance system for the visually impaired smart glasses, gloves, and shopping carts provide auditory and tactile feedback, IEEE CONSUMER ELECTRONICS MAGAZINE 6 (1) (2017) 73–81.

[50] A. Franco, D. Maltoni, S. Papi, Grocery product detection and recognition, Expert Systems with Applications 81 (2017) 163–176.

[51] K. A. Thakoor, S. Marat, P. J. Nasiatka, B. P. McIntosh, F. E. Sahin, A. R. Tanguay, J. D. Weiland, L. Itti, Attention biased speeded up robust features (ab-surf): a neurally-inspired object recognition algorithm for a wearable aid for the visually impaired, in: Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on, IEEE, 2013, pp. 1–6.

[52] A. Ray, N. Kumar, A. Shaw, D. Prasad Mukherjee, U-pc: Unsupervised planogram compliance, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 586–600.

[53] K. Iwamoto, R. Mase, T. Nomura, Bright: A scalable and compact binary descriptor for low-latency and high accuracy object identification, in: Image Processing (ICIP), 2013 20th IEEE International Conference on, IEEE, 2013, pp. 2915–2919.

[54] K. Higa, K. Iwamoto, T. Nomura, Multiple object identification using grid voting of object center estimated from keypoint matches, in: Image Processing (ICIP), 2013 20th IEEE International Conference on, IEEE, 2013, pp. 2973–2977.

[55] E. R. Dougherty, An introduction to morphological image processing, Tutorial texts in optical engineering.

[56] E. Frontoni, M. Contigiani, G. Ribighini, A heuristic approach to evaluate occurrences of products for the planogram maintenance, in: Mechatronic and Embedded Systems and Applications (MESA), 2014 IEEE/ASME 10th International Conference on, IEEE, 2014, pp. 1–6.

[57] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, Vol. 1, IEEE, 2005, pp. 886–893.

[58] M. Marder, S. Harary, A. Ribak, Y. Tzur, S. Alpert, A. Tzadok, Using image analytics to monitor retail store shelves, IBM Journal of Research and Development 59 (2/3) (2015) 3–1.

[59] G. Varol, R. S. Kuzu, Toward retail product recognition on grocery shelves, in: Sixth International Conference on Graphic and Image Processing (ICGIP 2014), International Society for Optics and Photonics, 2015, pp. 944309–944309.

[60] I. Sobel, History and definition of the sobel operator, Retrieved from the World Wide Web.

[61] J. Canny, A computational approach to edge detection, in: Readings in Computer Vision, Elsevier, 1987, pp. 184–203.

[62] C. P. Papageorgiou, M. Oren, T. Poggio, A general framework for object detection, in: Computer vision, 1998. sixth international conference on, IEEE, 1998, pp. 555–562.

[63] J. Liu, Y. Liu, Grasp recurring patterns from a single view, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2003–2010.

[64] S. Liu, H. Tian, Planogram compliance checking using recurring patterns, in: Multimedia (ISM), 2015 IEEE International Symposium on, IEEE, 2015, pp. 27–32.

[65] S. Liu, W. Li, S. Davis, C. Ritz, H. Tian, Planogram compliance checking based on detection of recurring patterns, IEEE MultiMedia 23 (2) (2016) 54–63.

[66] E. Goldman, J. Goldberger, Large-scale classification of structured image classification from conditional random field with deep class embedding, CoRR abs/1705.07420. arXiv:1705.07420.

[67] C. L. Novak, S. A. Shafer, Anatomy of a color histogram, in: Computer Vision and Pattern Recognition, 1992. Proceedings CVPR'92., 1992 IEEE Computer Society Conference on, IEEE, 1992, pp. 599–605.

[68] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Transactions on pattern analysis and machine intelligence 20 (11) (1998) 1254–1259.

[69] D. Jameson, L. M. Hurvich, Essay concerning color constancy, Annual review of psychology 40 (1) (1989) 1–24.

[70] T. Gevers, A. W. Smeulders, Color-based object recognition, Pattern recognition 32 (3) (1999) 453–464.

[71] T. Gevers, A. W. Smeulders, Content-based image retrieval by viewpoint-invariant color indexing, Image and vision computing 17 (7) (1999) 475–488.

[72] T. Gevers, A. W. Smeulders, Pictoseek: Combining color and shape invariant features for image retrieval, IEEE transactions on Image Processing 9 (1) (2000) 102–119.

[73] T. Gevers, Robust histogram construction from color invariants, in: Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference On, Vol. 1, IEEE, 2001, pp. 615–620.

[74] A. Diplaros, T. Gevers, I. Patras, et al., Color-shape context for object recognition, in: IEEE Workshop on Color and Photometric Methods in Computer Vision, 2003, pp. 1–8.

[75] T. Gevers, H. Stokman, Robust histogram construction from color invariants for object recognition, IEEE transactions on pattern analysis and machine intelligence 26 (1) (2004) 113–118.

[76] A. Diplaros, T. Gevers, I. Patras, Combining color and shape information for illumination-viewpoint invariant object recognition, IEEE Transactions on Image Processing 15 (1) (2006) 1–11.

[77] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: Convolutional architecture for fast feature embedding, in: Proceedings of the 22nd

ACM international conference on Multimedia, ACM, 2014, pp. 675–678.

[78] A. Dingli, I. Mercieca, Multimedia interfaces for people visually impaired, in: Advances in Design for Inclusion, Springer, 2016, pp. 487–495.

[79] P. Jund, N. Abdo, A. Eitel, W. Burgard, The freiburg groceries dataset, arXiv preprint arXiv:1611.05799.

[80] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

[81] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.

[82] T. Chong, I. Bustan, M. Wee, Deep learning approach to planogram compliance in retail stores.

[83] K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman, Return of the devil in the details: Delving deep into convolutional nets, arXiv preprint arXiv:1405.3531.

[84] L. Karlinsky, J. Shtok, Y. Tzur, A. Tzadok, Fine-grained recognition of thousands of object categories with single-example training, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4113–4122.

[85] Y. A. LeCun, L. Bottou, G. B. Orr, K.-R. Müller, Efficient backprop, in: Neural networks: Tricks of the trade, Springer, 2012, pp. 9–48.

[86] Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives, IEEE transactions on pattern analysis and machine intelligence 35 (8) (2013) 1798–1828.

[87] A. M. Treisman, G. Gelade, A feature-integration theory of attention, Cognitive psychology 12 (1) (1980) 97–136.

[88] N. Bruce, J. Tsotsos, An information theoretic model of saliency and visual search, Attention in cognitive systems. Theories and systems from an interdisciplinary viewpoint (2007) 171–183.

[89] M. George, D. Mircic, G. Soros, C. Floerkemeier, F. Mattern, Fine-grained product class recognition for assisted shopping, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015, pp. 154–162.

[90] S. Advani, B. Smith, Y. Tanabe, K. Irick, M. Cotter, J. Sampson, V. Narayanan, Visual co-occurrence network: using context for large-scale object recognition in retail, in: Embedded Systems For Real-time Multimedia (ESTIMedia), 2015 13th IEEE Symposium on, IEEE, 2015, pp. 1–10.

[91] R. Szeliski, Computer vision: algorithms and applications, Springer Science & Business Media, 2010.

[92] J. Schwiegerling, Field guide to visual and ophthalmic optics, SPIE, 2004.

[93] M. Fritz, B. Leibe, B. Caputo, B. Schiele, Integrating representative and discriminant models for object category detection, in: Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, Vol. 2, IEEE, 2005, pp. 1363–1370.

[94] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: Computer vision and pattern recognition, 2006 IEEE computer society conference on, Vol. 2, IEEE, 2006, pp. 2169–2178.

[95] R. O. Duda, P. E. Hart, Use of the hough transformation to detect lines and curves in pictures, Communications of the ACM 15 (1) (1972) 11–15.

[96] A. L. Kesidis, N. Papamarkos, On the gray-scale inverse hough transform, Image and Vision Computing 18 (8) (2000) 607–618.

[97] K. Mikolajczyk, C. Schmid, Scale & affine invariant interest point detectors, International journal of computer vision 60 (1) (2004) 63–86.

[98] R. O. Duda, P. E. Hart, D. G. Stork, Pattern classification, John Wiley & Sons, 2012.

[99] P. Piccinini, A. Prati, R. Cucchiara, Real-time object detection and localization with sift-based clustering, Image and Vision Computing 30 (8) (2012) 573–587.

[100] S. Zickler, M. M. Veloso, Detection and localization of multiple objects, in: Humanoid Robots, 2006 6th IEEE-RAS International Conference on, IEEE, 2006, pp. 20–25.

[101] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, Y. Gong, Locality-constrained linear coding for image classification, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE, 2010, pp. 3360–3367.

[102] B. Yao, A. Khosla, L. Fei-Fei, Combining randomization and discrimination for fine-grained image categorization, in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE, 2011, pp. 1577–1584.

[103] J. Kim, C. Liu, F. Sha, K. Grauman, Deformable spatial pyramid matching for fast dense correspondences, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2307–2314.

[104] D. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley, 1989.

[105] M. Jaderberg, A. Vedaldi, A. Zisserman, Deep features for text spotting, in: European conference on computer vision, Springer, 2014, pp. 512–528.

[106] R. Smith, An overview of the tesseract ocr engine, in: Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on, Vol. 2, IEEE, 2007, pp. 629–633.

[107] S. Singh, A. Gupta, A. Efros, Unsupervised discovery of mid-level discriminative patches, Computer Vision–ECCV 2012 (2012) 73–86.

[108] D. J. MacKay, Information-based objective functions for active data selection, Neural computation 4 (4) (1992) 590–604.

[109] J. H. Friedman, J. L. Bentley, R. A. Finkel, An algorithm for finding best matches in logarithmic expected time, ACM Transactions on Mathematical Software (TOMS) 3 (3) (1977) 209–226.

[110] A. Gionis, P. Indyk, R. Motwani, et al., Similarity search in high dimensions via hashing, in: VLDB, Vol. 99, 1999, pp. 518–529.

[111] J. S. Beis, D. G. Lowe, Shape indexing using approximate nearest-neighbour search in high-dimensional spaces, in: Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on, IEEE, 1997, pp. 1000–1006.

[112] M. A. Fischler, R. C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Communications of the ACM 24 (6) (1981) 381–395.

[113] B. Leibe, A. Leonardis, B. Schiele, Robust object detection with interleaved categorization and segmentation, International journal of computer vision 77 (1-3) (2008) 259–289.

[114] M. Muja, D. G. Lowe, Fast approximate nearest neighbors with automatic algorithm configuration., VISAPP (1) 2 (331-340) (2009) 2.

[115] M. Muja, D. G. Lowe, Fast matching of binary features, in: Computer and Robot Vision (CRV), 2012 Ninth Conference on, IEEE, 2012, pp. 404–410.

[116] D. Nister, H. Stewenius, Scalable recognition with a vocabulary tree, in: Computer vision and pattern recognition, 2006 IEEE computer society conference on, Vol. 2, Ieee, 2006, pp. 2161–2168.

[117] J. Wang, J. Xiao, W. Lin, C. Luo, Discriminative and generative vocabulary tree: with application to vein image authentication and recognition, Image and Vision Computing 34 (2015) 51–62.

[118] B. Matei, Y. Shan, H. S. Sawhney, Y. Tan, R. Kumar, D. Huber, M. Hebert, Rapid object indexing using locality sensitive hashing and joint 3d-signature space estimation, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (7) (2006) 1111–1126.

[119] B. Kulis, K. Grauman, Kernelized locality-sensitive hashing for scalable image search, in: Computer Vision, 2009 IEEE 12th International Conference on, IEEE, 2009, pp. 2130–2137.

[120] J. Wang, S. Kumar, S.-F. Chang, Semi-supervised hashing

for scalable image retrieval, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE, 2010, pp. 3424–3431.

[121] A. Andoni, P. Indyk, Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions, in: Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on, IEEE, 2006, pp. 459–468.

[122] B. W. Silverman, Density estimation for statistics and data analysis, Vol. 26, CRC press, 1986.

[123] S. Choudhary, A. J. Trevor, H. I. Christensen, F. Dellaert, Slam with object discovery, modeling and mapping, in: Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on, IEEE, 2014, pp. 1018–1025.

[124] O. Boiman, E. Shechtman, M. Irani, In defense of nearest-neighbor based image classification, in: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE, 2008, pp. 1–8.

[125] I. Jolliffe, Principal component analysis, Wiley Online Library, 2002.

[126] R. Brunelli, Template matching techniques in computer vision: theory and practice, John Wiley & Sons, 2009.

[127] C. Harris, M. Stephens, A combined corner and edge detector., in: Alvey vision conference, Vol. 15, Manchester, UK, 1988, pp. 10–5244.

[128] D. Chai, K. N. Ngan, Face segmentation using skin-color map in videophone applications, IEEE Transactions on circuits and systems for video technology 9 (4) (1999) 551–564.

[129] J. Zhang, M. Marszalek, S. Lazebnik, C. Schmid, Local features and kernels for classification of texture and object categories: A comprehensive study, in: Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on, IEEE, 2006, pp. 13–13.

[130] C. Wang, K. Huang, How to use bag-of-words model better for image classification, Image and Vision Computing 38 (2015) 65–74.

[131] K. Fukunaga, Statistical pattern recognition, in: Handbook Of Pattern Recognition And Computer Vision, World Scientific, 1999, pp. 33–60.

[132] J. D. Lafferty, A. McCallum, F. C. N. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in: Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, Morgan Kaufmann Publishers Inc., 2001, pp. 282–289.

[133] P. A. Devijver, Baum's forward-backward algorithm revisited, Pattern Recognition Letters 3 (6) (1985) 369–373.

[134] G. D. Forney, The viterbi algorithm, Proceedings of the IEEE 61 (3) (1973) 268–278.

[135] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.

[136] R. Girshick, Fast r-cnn, arXiv preprint arXiv:1504.08083.

[137] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in neural information processing systems, 2015, pp. 91–99.

[138] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Computer Vision (ICCV), 2017 IEEE International Conference on, IEEE, 2017, pp. 2980–2988.

[139] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg, Ssd: Single shot multibox detector, in: European conference on computer vision, Springer, 2016, pp. 21–37.

[140] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, Vol. 1, IEEE, 2001, pp. I–I.

[141] C. Papageorgiou, T. Poggio, Trainable pedestrian detection, in: Image Processing, 1999. ICIP 99. Proceedings. 1999 International Conference on, Vol. 4, IEEE, 1999, pp. 35–39.

[142] Y. Freund, R. Schapire, N. Abe, A short introduction to boosting, Journal-Japanese Society For Artificial Intelligence 14 (771-780) (1999) 1612.

[143] C. Cortes, V. Vapnik, Support-vector networks, Machine learning 20 (3) (1995) 273–297.

[144] P. A. Viola, M. J. Jones, Object recognition system, uS Patent 7,031,499 (Apr. 18 2006).

[145] J. Shi, et al., Good features to track, in: Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on, IEEE, 1994, pp. 593–600.

[146] M. Leordeanu, M. Hebert, A spectral technique for correspondence problems using pairwise constraints, in: Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, Vol. 2, IEEE, 2005, pp. 1482–1489.

[147] L. Rabiner, B. Juang, An introduction to hidden markov models, ieee assp magazine 3 (1) (1986) 4–16.

[148] X. He, R. S. Zemel, M. Á. Carreira-Perpiñán, Multiscale conditional random fields for image labeling, in: Computer vision and pattern recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE computer society conference on, Vol. 2, IEEE, 2004, pp. II–II.

[149] A. Andreopoulos, J. K. Tsotsos, 50 years of object recognition: Directions forward, Computer Vision and Image Understanding 117 (8) (2013) 827–891.

[150] D. Cremers, Optimal solutions for semantic image decomposition, Image and Vision Computing 30 (8) (2012) 476–477.

[151] J. Yao, S. Fidler, R. Urtasun, Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, 2012, pp. 702–709.

[152] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, A. Yuille, The role of context for object detection and semantic segmentation in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 891–898.

[153] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, A. Yuille, Adversarial examples for semantic segmentation and object detection, in: International Conference on Computer Vision. IEEE, 2017.

[154] L. Cruz, D. Lucio, L. Velho, Kinect and rgbd images: Challenges and applications, in: 2012 25th SIBGRAPI Conference on Graphics, Patterns and Images Tutorials, IEEE, 2012, pp. 36–49.