U-PC: Unsupervised Planogram Compliance

Archan Ray[1], Nishant Kumar[2], Avishek Shaw[3], and Dipti Prasad Mukherjee[4]

[1] University of Massachusetts Amherst, MA 01003, USA, ray@cs.umass.edu
[2] SMART-FM, Singapore, nishant@smart.mit.edu
[3] TCS Limited, India, shaw.avishek@tcs.com
[4] Indian Statistical Institute, Kolkata 700108, India, dipti@isical.ac.in

Abstract. We present an end-to-end solution for recognizing merchandise displayed in the shelves of a supermarket. Given images of individual products, which are taken under ideal illumination for product marketing, the challenge is to find these products automatically in the images of the shelves. Note that the images of shelves are taken using hand-held camera under store level illumination. We provide a two-layer hypotheses generation and verification model. In the first layer, the model predicts a set of candidate merchandise at a specific location of the shelf while in the second layer, the hypothesis is verified by a novel graph theoretic approach. The performance of the proposed approach on two publicly available datasets is better than the competing approaches by at least 10%.

Keywords: Planogram Compliance · Merchandise Recognition

1 Introduction

The display of merchandise on the shelves of a retail store follows a specific strategy, commonly known as planogram. To re-conciliate a planogram, an inspector visits each shelf and manually checks the availability of the merchandise as specified in the planogram. This is an expensive and error-prone exercise. We propose to capture the images of these shelves using hand held camera and provide an end-to-end solution to detect the products available on the shelves from their images. We expect that our tool may be used for planogram compliance. We do not impose any restriction on the camera type and the store lighting condition for wider acceptability of our proposal.

We assume that individual product images, typically used for marketing, are available in an image dataset. In addition we assume that the physical dimensions of the shelves and the individual products are available in any unit of length. We also assume that the planogram is not available to our software. In other words, we do not have any prior information about the location of products on the shelves. Therefore, for our problem, object recognition and localization are equally important.

A typical shelf image is shown in Fig. 1(a). Individual images of products present on the shelf and available in the dataset of product images are shown

Fig. 1. (a) An example shelf image. (b)-(e) Sample product images. (f)-(g) Poor quality product images cropped from shelf image. (d)-(e) and (f)-(g) are images of same products, respectively.

in Fig. 1(b) to (d). Notice the variation in illumination, resolution and quality of images between Fig. 1(a) and (b) to (d). Also note that dataset may contain product images (Fig. 1(d) and (e)) which are not available on the shelf.

In this paper we address recognition and localization of multiple objects in a scene at one go. The approaches in [7] propose a set of view-invariant transformations of *rgb* color vector for recognition of consumer products on the shelf. These transformations cannot handle the differences in resolution and illumination between the shelf image and the product image. Further, the unstable imaging conditions may arise from specular reflections, from shiny packages of products, instability in taking images by the shelf inspector. Typical degradation of the quality of product images cropped from a shelf image are shown in Fig. 1(f) and (g).

Zhang *et al.* [17] extract SIFT like features from a region using Harris-Affine interest region detector [13]. Both product image and shelf image are divided into sub-images and matched using histogram of features. We have compared our work with [17]. A combination of SIFT and histogram based matching is used for identifying grocery products in [12].

In [5], 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) [15] model using dense SIFT features of product images present in the dataset. A discriminative random forest [16] is then trained with LLC features of product images. Using the trained model, a multi-class ranking of products is estimated at a location by classifying each block of the shelf image.

The authors in [10] perform a deformable spatial pyramid based fast dense pixel matching and genetic algorithm based optimization scheme [8] for localization and recognition of products in the shelf image. In a variation of the above approach in [6], the authors have also integrated text based recognition [9] and features derived from discriminative patches as in [14]. The products are recognized using SVM; the recognition performance is improved using active learning [11] through user feedback.

The approach in [5] looks at the object localization problem more as an image retrieval challenge. Therefore, [5] fails to serve a key challenge that we are addressing is that of simultaneous detection, recognition and localization of multiple products. To explain the challenge further, assume products a and b are neighbors in a shelf. Assume a is identified incorrectly as another product c. The

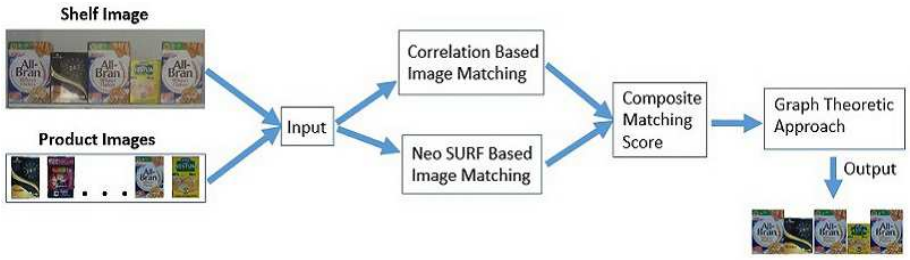


Fig. 2. Overall block diagram of the proposed scheme.

width of c is different from a. Then the region to be cropped to identify b will be incorrect. This is an important issue for retailers who want to locate both a and b instead of retrieving them in isolation. We have handled this problem by introducing a novel graph-theoretic approach to locate all products at the same time.

As mentioned earlier, we have introduced a two-layer approach of hypothesis generation and verification for localization and recognition of multiple objects on the shelf. At the initial stage, our strategy is to allow for exhaustive match of product images with the content of shelf image. This we refer as hypotheses generation. From the set of hypotheses, multiple products are predicted at a particular location on the shelf. In the second layer, one particular product out of these multiple predictions at a particular location is selected based on a graph theoretic approach. The overall block diagram of the proposed scheme is presented in Fig. 2. In the next section, we present the proposed matching scheme. The results are discussed in Section 3 followed by conclusions.

2 Image Matching

Multiple product identification in the wild is difficult due to many variabilities. These variables include the unknown scale of the products in the shelf image and color variability due to unconstrained illumination. Images taken from two different brands of cameras may result in variation of color space of images [2]. In a given row of multiple products, identification of a product at a particular location influences identification of neighboring products in the row. We address these variables in our methodology as presented next.

2.1 Hypothesis Generation

Assume images of N products are available in the dataset of product images $\mathcal{D}$. Each such product image is referred as $\mathcal{D}_p$, $p = 1, 2, \cdots, N$. Typical examples of $\mathcal{D}_p$ are shown in Fig. 1(b) to (e). We assume physical dimensions of these individual products in any suitable unit of length are available. Note that different products in $\mathcal{D}_p$ come in different physical and pixel dimensions.

Let the shelf image where multiple products are present, be represented by I_s. An example I_s image is shown in Fig. 1(a). We do not put any restriction on the camera used to capture I_s. But we do have the physical dimension of the shelf in any suitable unit of length. The problem is to find the location of $\mathcal{D}_p$ in I_s.

Dimensions of each image in $\mathcal{D}$ in pixels can be converted to the length-to-pixel scale of I_s. Let the rescaled dataset of product images be $\mathcal{D}'$. To emphasize, pixel dimensions of the pth product, $\mathcal{D}'_p$ is an approximation and does not represent true scale of the product image $\mathcal{D}_p$ in I_s. And this is one of the major challenges of the proposed problem.

Let there be c columns in I_s. For every location i, $i = 1, 2, \cdots, c$, of I_s, N number of images are cropped from I_s. The dimension of each of N crops is same as that of $\mathcal{D}'_p$, $p = 1, 2, \cdots, N$. Each of the N crops are correlated with the corresponding image of $\mathcal{D}'_p$ to calculate the Pearson correlation coefficient. Correlation is done in three separate Lab channels and the average of three correlation coefficients are computed. Therefore, at every location i, $i = 1, 2, \cdots, c$, of I_s, there are N correlation scores.

As mentioned earlier, scaling of $\mathcal{D}_p$ to $\mathcal{D}'_p$ is an approximation. To counter this, we isotropically vary (upscale and downscale) the dimensions of $\mathcal{D}'_p$ within a range $[-l/2, l/2]$ where $l \in \mathbb{R}$. Due to these additional resizes of $\mathcal{D}'_p$, additional l number of transformed images of $\mathcal{D}'_p$ are cropped at location i of I_s.

Combining above two scenarios, at a particular location i, $i = 1, 2, \cdots, c$, of I_s, $(N + N \times l)$ number of crops of I_s equivalent to image size $\mathcal{D}'_p$ and its scaled version are correlated with $\mathcal{D}'_p$. Algorithm 1 sums up this proposed hypothesis generation scheme. Function `crop` crops a patch at ith column of I_s equivalent to

Algorithm 1 Algorithm for hypothesis generation

procedure HYPOTHESIS GENERATION($\mathcal{D}', I_s$)
  Define $l \in \mathbb{R}^+$
  $[h, c] = \text{size}(I_s)$
  Initialize $C_r = \text{zeros}(c, N)$
  for $i \in [1, c]$; $p \in [1, N]$; **do**
    $Q = \text{crop}(I_s, \text{size}(\mathcal{D}'_p), i)$
    $C_r[i, p] += \text{match\_score}(Q, \mathcal{D}'_p)$
  end for
  for $i \in [1, c]$; $p \in [1, N]$; $k \in [-l/2, l/2]$ **do**
    $Q = \text{crop}(I_s, \text{size}(\mathcal{D}'_p), i, k)$
    $C_r[i, p] += \text{match\_score}(Q, \mathcal{D}'_p)$
  end for
  return C_r
end procedure

the size of $\mathcal{D}'_p$. In an overloaded version of `crop`, the k times scaled $\mathcal{D}'_p$ is cropped at ith location. The `match_score` function computes average of correlation coefficients in Lab channels between the cropped patch and $\mathcal{D}'_p$ or its scaled version.

The cumulative score C_r for pth product at ith location of I_s are cumulated for all possible $(N + N \times l)$ values. The top m cumulative scores C_r represent m possible products likely to be present at column i of I_s, $i = 1, 2, \cdots, c$. We refer these m products as top m matches at ith location.

2.2 Matching Strategies

We have explored several image matching strategies between Q and $\mathcal{D}'_p$ or its scaled version other than straightforward correlation. However, we have not seen significant difference in our desired result due to matching strategies. Again, it is not our intention to get an exact match at this stage but to keep the right product in the list of top m matches. The straightforward correlation based matching proposed in this paper suffices as it always keeps the right product at a particular location within top m matches.

We have observed that complicated matching strategies with a number of tunable parameters do not give any advantage in the choice of the top m possible products at ith location. We have seen that $m = 3$ consistently finds the correct product those are likely to be present at location i. These top m matches are the hypotheses at a particular column location.

Transformation of $\mathcal{D}$ to $\mathcal{D}'$ under store level illumination is a difficult proposition. While the height of the shelf in any unit of length is known, it is impossible to find the exact boundary of the shelf in a shelf image taken by a hand held camera. However, the height of any product cannot be more than the height of the shelf. Therefore, to take care of all possibilities, height of $\mathcal{D}'_p$ is scaled up to the height of the shelf maintaining the aspect ratio of $\mathcal{D}'_p$. This upscaling determines the value of $l/2$. By the same amount $\mathcal{D}'_p$ is downscaled $l/2$ times. This determines the choice of l.

A test suite is designed with 19 product images of $\mathcal{D}$ to calculate C_r for identifying products in Fig. 1(a). C_r values for 19 products at three random locations of the shelf of Fig. 1(a) are shown in Fig. 3. The histogram in Fig. 3 shows that product number 12, 19 and 3 are the top-3 likely candidates based on cumulative score C_r at column 2 of Fig. 1(a). Similarly, for column 32, the likely products are 7, 3 and 12.

This concludes our hypothesis generation step. Next we find the shape based matching scores for top m possible products using SURF [1].

2.3 Neo SURF based Matching

The Neo SURF (NSURF) introduced for our problem is a speeded up customized version of SURF. The primary motivation of using SURF is to complement our intensity based correlation score with the matching using shape based features. SURF had been used earlier successfully in similar problems [5]. NSURF differs from SURF in two aspects. First, in our typical use-cases, rotation invariance is not required [1]. The slant of the box on a shelf with respect to its upright position is typically $\pm 15^\circ$. Such minor variations in slant do not affect the estimation of key points with SURF. Second, for very small images, the performance of

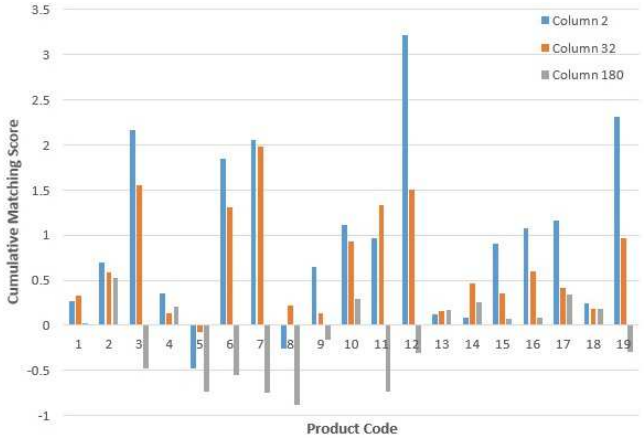


Fig. 3. C_r score of Fig. 1(a) for all 19 products at the 2nd, 32nd and 180th columns.

SURF is found to be poor. The larger sized kernels account for this. For our implementation, we have restricted the kernel size to a maximum of 51×51. Third, we have used only two lower octaves for NSURF.

NSURF is applied on the cropped patches corresponding to size of $\mathcal{D}'_p$ for top m products at the ith location of I_s. Let the number of keypoints obtained using NSURF on the patch of I_s and $\mathcal{D}'_p$ be $k1$ and $k2$ respectively. Let the feature vector obtained from each keypoints be K_x, $1 \leq x \leq k1$ and K'_y, $1 \leq y \leq k2$, respectively. Each K_x or K'_y is identified by 64 dimensional vector. Let any particular K_x (say, K_d) match best with any particular K'_y (say K'_e) with an Euclidean distance θ. For our implementation, for a potential match between K_d and K'_e, we have chosen a conservative threshold as $\theta \leq 0.04$. In addition we have ensured that the ratio between minimum distance θ and the second minimum distance of K_d from all other K'_y except K'_e should be less than 0.4. This ensures a dominant yet reliable matching of keypoints between the patch of I_s (corresponding to top m products) and $\mathcal{D}'_p$. The total number of such matched keypoints are taken as NSURF score. Next we present the strategy to combine the correlation score derived in Section 2.1 with NSURF score.

2.4 Combining Correlation and NSURF Scores

The NSURF score, say U, is significantly higher compared to cumulative correlation score C_r. We design a composite score with a motivation that magnifies the discrimination between top m products at ith location. The designed composite score C_s is given by

$$C_s = U^{C_r}. \tag{1}$$

Similar types of products (for example, breakfast cereals or milk containers), similar in terms of dimension but dissimilar in packaging, are usually available in one given shelf. The packages are rich in content generating a number of

NSURF keypoints. Therefore, $U \in \mathbb{Z}^+$ has higher and better discriminatory value compared to C_r, which for an ideal match should be close for similar products. If we raise U to C_r, the value is $\in \mathbb{R}^+$. In other words, (1) helps in magnifying the difference between products and leads to a positive value.

Combining algorithm 1 and (1), at each column of the shelf image, we have composite scores for top m products. The next task is to select the winning product out of these top m possible products. This choice of winner product should consider all columns in the shelf simultaneously. We employ a directed graph for this purpose which is detailed next.

2.5 Construction of Directed Graph

Given that there are m possible products at a particular column of I_s, one straightforward approach could be to pick up the product with highest C_s. However, the product with highest C_s may not be the correct product. Further if a product $\mathcal{D}'_p$ is chosen at ith column of I_s, no other product should be selected for the width of $\mathcal{D}'_p$. However, within the width of $\mathcal{D}'_p$, there exists other products whose C_s value may be higher than the $\mathcal{D}'_p$ chosen at the ith column. Therefore all possible column locations of I_s should be considered simultaneously in order to find a winner product at the ith column of I_s.

To allow for top m products at all positions of I_s to compete with each other with their C_s score, we construct a DAG, directed acyclic graph, $\mathcal{G}(V, E)$. An arbitrary source node $\mathcal{S}$ and a sink node $\mathcal{T}$ is added to $\mathcal{G}$. Therefore, $\mathcal{G}(V, E)$ has total $(cm + 2)$ nodes. The edges are defined in the matrix E. All nodes in $\mathcal{G} - \mathcal{S}$ have an incoming edge from $\mathcal{S}$ such that $E[\mathcal{S}, \{\mathcal{G} - \mathcal{S}\}] = \epsilon$. Similarly, all nodes have directed edges to $\mathcal{T}$ such that $E[\{\mathcal{G} - \mathcal{T}\}, \mathcal{T}] = \epsilon$. A node $v_{ij} \in V$ represents jth product, $1 \leq j \leq m$ at ith location, $1 \leq i \leq c$.

For any node $v_{ij}, v_{op} \in \mathcal{G} - \{\mathcal{S}, \mathcal{T}\}$, $E[v_{ij}, v_{op}] = C_s[i, j]$, iff $(o - i) \geq width(\mathcal{D}'_j)$. The width of jth product is $width(\mathcal{D}'_j)$. In other words, there is an edge from one product $\mathcal{D}'_j$ to another product $\mathcal{D}'_p$ weighing equivalent to the composite score of $\mathcal{D}'_j$ at position i iff $\mathcal{D}'_p$ is at least as far as the width of $\mathcal{D}'_j$.

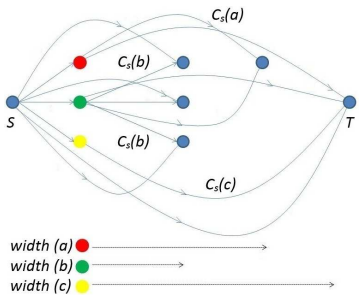


Fig. 4. An example graph $\mathcal{G}$ is shown. The top 3 possible products at the 1st column are a, b and c with composite scores $C_s(a)$, $C_s(b)$ and $C_s(c)$ and widths as $width(a)$, $width(b)$ and $width(c)$, respectively.

A typical $\mathcal{G}$ with some example edge weights is shown in Fig. 4. The red, green and the yellow dots represent products at the first column. These dots represent top m products ($m = 3$ in this case) in terms of score C_s. The black straight lines at the bottom show the width of each product.

We obtain the maximum weighted path in this graph. We expect that the sum of the edge weights or composite scores to be maximum in case the products are identified correctly considering all possible columns of the shelf. Any incorrect placement of the product(s) should lead to the sum of composite scores being lower than that due to correct placement. Obtaining a maximum weighted path in any graph is an NP-hard problem. Since this is a directed acyclic graph, we have negated the edge weights and obtained the minimum weighted path using Bellman-Ford algorithm [3]. Next we justify our choice of maximum weighted path.

2.6 Justification for using Maximum Weighted Path

In this section, we prove the following statement: *Solving for* detection of multiple products in a shelf image in I_s (problem A) *is equivalent to solving for* maximum weighted path in a graph $\mathcal{G}$ (problem B).

Proof. We begin by analyzing the complexity of the graph construction from I_s introduced in Section 2.5. Let n be the number of nodes in the graph. The algorithm 1 is $\mathcal{O}(n^4)$ and the process of calculation of composite score is $\mathcal{O}(n)$. The construction of the graph is a $\mathcal{O}(n)$ algorithm. Thus the overall construction of the graph $\mathcal{G}$ is polynomial in n. We only need to show that if there exists a solution in B, then there will exist a solution in A. For this we first need to understand what is a solution in A.

Define *ideal crop* Q_p of the size of a product $\mathcal{D}'_p$ from I_s as the perfect crop of one instance of $\mathcal{D}'_p$ in I_s. Let product $\mathcal{D}'_p$ appear at the oth location of I_s. Naturally, the composite score for Q_p from I_s would be maximum at column o for the product $\mathcal{D}'_p$.

Now consider another product $\mathcal{D}'_j$ be present in I_s, and its ideal patch be Q_j. Naturally the ideal crop for $\mathcal{D}'_j$ will not be in the range $[o, width(Q_p) + o)$. Without any loss of generalization, let the original position of Q_j be at column i, such that $i > o$. The nodes representing products for Q_p and Q_j in $\mathcal{G}$ will be connected by a weighted directed edge from $v_{op} \rightarrow v_{ij}$. Considering the construction of the graph $\mathcal{G}$ detailed in Section 2.5, we need to show that the maximum weighted path will have to go through the two nodes v_{op} and v_{ij}.

Let $\mathcal{P}$ be the maximum weighted path in $\mathcal{G}$ that does not go through $\{v_{op}, v_{ij}\}$. Thus there exist other nodes in the neighborhood of $\{v_{op}, v_{ij}\}$ which has a higher composite score. Let (α, β) be the NSURF scores of the two patches Q_p and Q_j respectively. Since these are ideal patches the correlation score is 1 for both. Thus the composite score for the path through only these two nodes is $(\alpha + \beta)$. It is obvious that,

$$(\alpha + \beta) \geq 2[\min(\alpha, \beta)]. \tag{2}$$

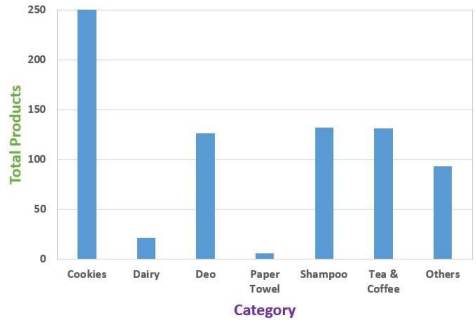


Fig. 5. Categories of product images in the in-house dataset.

Let the expected value of the NSURF score of falsely matched products around the region $\{o, i\}$ be at most γ. Let ζ be the expected correlation score of incorrectly matched products. Let there be κ such products that can be fit in the region $\{o, i\}$. Then the weight of the path is $\kappa\gamma^{\zeta}$. Without loss of generality, we know that $\zeta << 1$ and $\gamma << \min(\alpha, \beta)$. Thus $\gamma^{\zeta} << \min(\alpha, \beta)$.

Given the information above, we would now like to analyze the variation of the total weight of the path $\mathcal{P}$ with respect to the width of each product. Let w be the width of I_s; w_1, w_2 be the minimum and maximum widths of any product in $\mathcal{D}'$ respectively. Thus $\kappa \in [\frac{w}{w_2}, \frac{w}{w_1}]$ and $\kappa \in \mathbb{Z}$. For our assumption to be false, we need to show $\kappa\gamma^{\zeta} \geq 2\min(\alpha, \beta)$. For the maximum value of κ, we have,

$$\frac{w}{w_1}\gamma^{\zeta} \geq 2[\min(\alpha, \beta)]. \tag{3}$$

Without loss of generality we can assume that the maximum matched NSURF points, α, are uniformly distributed in the space of an extracted crop Q. We do not assume other distributions because if we are able to show it works in uniform distribution, it can be implied that this would work for other distributions as well. Thus in an ideal case there is a quadratic relationship between an increase in width of Q and the number of matched keypoints for a particular product $\mathcal{D}'_p$. Let us assume that the number of matched keypoints increase at a sub-quadratic rate with the width for incorrect matches of Q. We know, $\frac{w}{w_1}$ decreases linearly with increasing w_1. But α increases with the width of Q, until the ideal crop dimensions are achieved and then becomes constant. Now, γ increases at a sub-quadratic rate, and ζ goes down as the correlation between the crop and the product goes down with the increasing width of the crop.

Since ζ is upper bounded by 1, the function $\frac{\alpha}{\gamma^{\zeta}}$ is constant for very low values of α, linear for ζ near 1, and increasing otherwise. Thus the $2[\min(\alpha, \beta)]$ always dominates $\frac{w}{w_1}\gamma^{\zeta}$. This contradicts (3). Thus any maximum path should pass through nodes in $\mathcal{G}$ represented by Q_p and Q_j. This concludes our proof that solving for problem B is equivalent to solving for problem A.

Since by construction our graph $\mathcal{G}$ is a DAG, we can convert the edges to negative weights and solve for the minimum weighted path. The minimum weighted

path using Bellman-Ford algorithm considering feed forward edge weights provides the final arrangement of products on the shelf. The result obtained using this proposed scheme is discussed next.

3 Results

The experiment is conducted both with in-house and publicly available datasets of product and shelf images. The in-house dataset consists of images of approximately 750 products in 7 categories. More than 2000 images of shelves are collected both from stores and lab settings. The category wise distributions of image dataset $\mathcal{D}$ are shown in Fig. 5. The proposed approach is also tested and compared using two publicly available datasets [17, 5]. We first show some qualitative results of the proposal followed by quantitative analysis.

The reconstructed image of Fig. 1(a), which is the output of DAG of Section 2.5 is shown in Fig. 6(a). Another example of reconstruction is shown in Fig. 6(b) where the top row is the shelf and the bottom row is the reconstructed shelf. The correct products are identified in spite of variation in illumination. Notice even minor variations (red stripe at the top of the box instead of green stripe) in the product labels of two consecutive boxes of Chocos cereals (at the right end of the shelf) could be recognized by our approach.

The bottom row of Fig. 7(a) shows yet another reconstruction of the original shelf image in the top row, where the first product, placed behind with respect to others, could not be recognized. Another reconstruction in Fig. 7(b) (bottom row) shows that even though Surf Excel bottle and Surf pouch has identical dominant texture on the front cover, the bottle and the pouch are identified correctly. The reconstruction of Fig. 7(c) bottom row clearly establishes the superiority of the proposal even under extreme specular reflection on some of the products.

Merler *et al.* [12] have proposed in situ product matching using color histogram. The result following [12] divides both $\mathcal{D}$ and I_s into smaller blocks and matches blocks using a score derived from intersection over union of areas under histograms. The result is shown in Fig. 8(a) whereas the output using

(a) (b)

Fig. 6. (a) Reconstruction of Fig. 1(a). (b) Correct reconstruction (bottom row) of shelf image (top row). Product with minor variation (two boxes at the right end) is correctly identified.

(a) (b) (c)

Fig. 7. (a) Failure case (bottom row) where the first product is incorrectly identified due to displaced position of the product in the shelf image in the top row. (b) Surf bottle and box having identical texture on the cover are identified correctly. (c) Correct reconstruction of the shelf image in spite of extreme specular reflection (top row: original shelf image, bottom row: reconstruction result).

proposed approach is shown in Fig. 8(b). Clearly, the accuracy of the proposed approach is better than matching using [12]. The role of integration of NSURF with correlation as opposed to selecting winner product at a location based only on maximum correlation score is shown in Fig. 9. The reconstruction result of Fig. 9(b) is better than that of Fig. 9(a). Similarly, NSURF alone cannot give desired result as opposed to the composite score as shown in Fig. 10. Note that all reconstructed results using proposed approach is the final output of DAG using composite score.

We start the quantitative analysis of our result by plotting the ROC of the reconstructed result. Assume there are N product images in our dataset where as r products are available in a given shelf. Typically, $r << N$. As mentioned earlier, we are solving both recognition and localization problem. If a product is identified at column i of the shelf and the algorithm predicts the product at a location $i \pm \delta$, we consider that the product is correctly identified. The shift δ is typically considered as 75mm for approximately 1000 mm wide shelf. Given this, True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) are defined as follows for each of the r products available in the shelf.

(a) (b)

Fig. 8. Result using (a) [12], (b) proposed approach (top row: original shelf image, bottom row: reconstruction).

(a) (b)

Fig. 9. (a) Winner product at a shelf location is selected based on maximum correlation score. (b) Result using composite score (top row: original shelf image, bottom row: reconstruction).

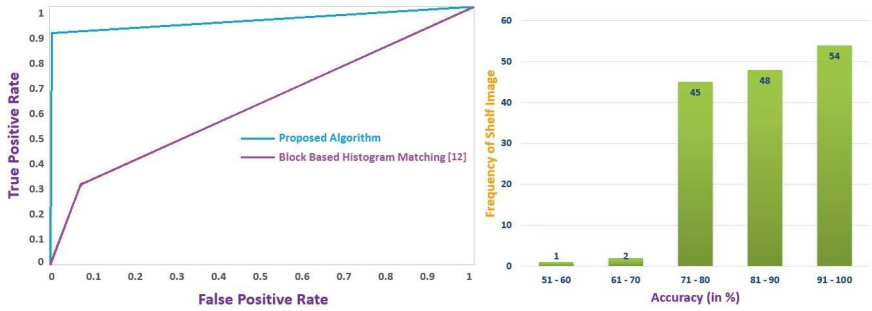
(a) (b)

Fig. 10. (a) Winner product at a shelf location is selected based on maximum NSURF score. (b) Result using composite score (top row: original shelf image, bottom row: reconstruction).

If product A is present at column i and the algorithm predicts A at column $i \pm \delta$, TP of product A is counted as 1. If a product other than A is present at column i and the algorithm predicts A at column $i \pm \delta$, FP of product A is counted as 1. If a product other than A is present at column i and the algorithm does not predict A at column $i \pm \delta$, TN of product A is counted as 1. If product A is present at column i and the algorithm predicts a product other than A at column $i \pm \delta$, FN of product A is counted as 1. The true positive rate (TP/(TP+FN)) versus false positive rate (FP/(FP+TN)) for 2000 shelf images is plotted in Fig. 11(a). The area under ROC for the proposed approach is significantly better compared to [12].

The proposed correlation and NSURF integrated graph based matching is applied on the entire set of shelf images. Approximately 2000 shelf images from lab and stores are organized in 150 racks, each rack containing multiple shelves. The histogram of accuracy values of rack-wise product identification using our approach is shown in Fig. 11(b). The accuracy value is the number of matches between the products of reconstructed result using our algorithm and the products in the ground truth divided by the total number of products present in the rack.

Additionally we have performed stress testing on 500 shelf images of Cookies category. Each of these shelf images is taken after varying camera angle within



(a) (b)

Fig. 11. (a) ROC Plot. (b) Accuracy using proposed approach (for example, reconstruction of 45 or 48 rack images has 71-80% or 81-90% accuracy, respectively).

$\pm 15°$ and at different camera-to-shelf distances. There are 9 unique products in these shelf images. Therefore, the product image dataset $\mathcal{D}$ initially contains 9 product images. The accuracy of detection of these product images in each of the 500 shelf images are calculated using both the proposed and block based histogram matching [12].

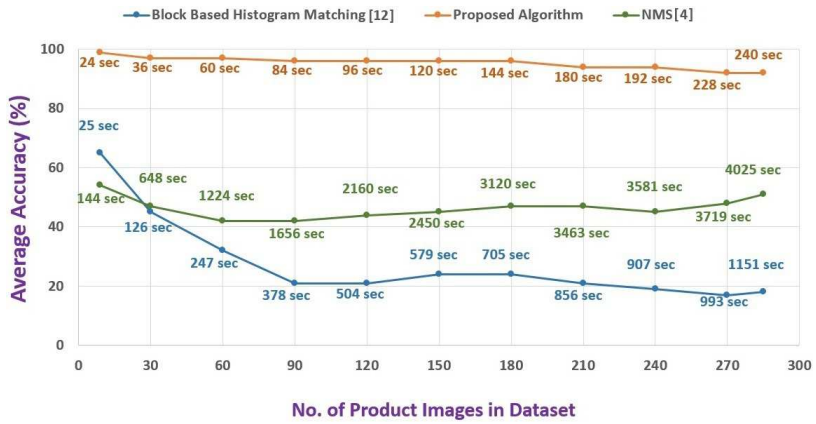


Fig. 12. Accuracy when the number of product images in $\mathcal{D}$ is increased from 9 to 285 in steps of 30. Corresponding computation times are mentioned.

For calculating accuracy, all the product images which are present in the shelf and identified correctly by both the algorithms are divided by the total number of products available in the shelves. The accuracy result averaged for 500 shelf images is reported as accuracy of product identification. The process is now repeated after increasing the size of the product image dataset $\mathcal{D}$ from 9 to 285

Table 1. Comparison of the proposed approach on publicly available and in-house datasets [17], [5].

	Inhouse	WebMarket [17]	Grocery [5]
Proposed	**92.4**	**90.8**	88.51
S1 [17]	41.6	62.03	51.28
S2 [17]	41.2	68.69	53.43
S3 [17]	36	59.29	63.05
MIC [5]	91.2	54.41	69.23
NMS [4]	47.56	51.02	**96.67**

in steps of 30. This experiment tests whether the proposed matching algorithm is confused with the additional 30 product images for each subsequent test. Note that these additional product images in multiples of 30 are anyway not present in the 500 shelf images under inspection. The accuracy plot against increasing size of dataset of product images is shown in Fig. 12. The corresponding computation time is also shown in Fig. 12. The experiment is repeated using NMS [4]. Fig. 12 shows that our proposal performs better compared to [4]. Further, the proposed image matching is not confused even with large number of spurious potential matches whereas [12] performs poorly with the increase in size of $\mathcal{D}$ containing product images not present in the shelf.

Finally the proposed approach is compared with two related approaches [17], [5]. Subsets of two publicly available datasets of [17], [5] along with in-house data are used for comparison. The approach in [17] has three matching functions $S1$, $S2$ and $S3$ as shown in Table 1. The key difference in our result with respect to competing approaches is that accuracy measure for the proposed approach on our dataset includes accuracy of both recognition and localization (module a shift of $\pm\delta$ from the exact location as mentioned earlier) of products on the shelf. The accuracy using the proposed approach is reported in the top row of the Table 1. For competing approaches [17], [5], the accuracy refers to recognition without any penalization for the inaccuracy of localization of the product.

4 Conclusions

We have provided an end-to-end solution for automatically recognizing products available on the shelf. No a priori information is used to preempt the type of products expected at a particular location of the shelf. Instead of looking at a particular discrete location of the shelf, all columns of the shelf are treated simultaneously using a novel graph based approach. We are now improving the approach by integrating single instance learning technique with the graph based search mechanism.

Acknowledgments

This work is partially supported by TCS Limited. The authors would like to thank Mr. Bikash Santra for his help in preparing the manuscript.

References

1. Bay, H., Tuytelaars, T., Van Gool, L.: Surf: Speeded up robust features. In: European conference on computer vision. pp. 404–417. Springer (2006)
2. Cheng, D., Prasad, D.K., Brown, M.S.: Illuminant estimation for color constancy: why spatial-domain methods work and the role of the color distribution. JOSA A **31**(5), 1049–1058 (2014)
3. Cormen, T.H., Leiserson, C.E., Rivest, R.L.: Introduction to Algorithms. Third Edition. MIT Press (2009)
4. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE transactions on pattern analysis and machine intelligence **32**(9), 1627–1645 (2010)
5. George, M., Floerkemeier, C.: Recognizing products: A per-exemplar multi-label image classification approach. In: European Conference on Computer Vision. pp. 440–455. Springer (2014)
6. George, M., Mircic, D., Soros, G., Floerkemeier, C., Mattern, F.: Fine-grained product class recognition for assisted shopping. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 154–162 (2015)
7. Gevers, T., Smeulders, A.W.: Color-based object recognition. Pattern recognition **32**(3), 453–464 (1999)
8. Goldberg, D.: Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley (1989)
9. Jaderberg, M., Vedaldi, A., Zisserman, A.: Deep features for text spotting. In: European conference on computer vision. pp. 512–528. Springer (2014)
10. Kim, J., Liu, C., Sha, F., Grauman, K.: Deformable spatial pyramid matching for fast dense correspondences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2307–2314 (2013)
11. MacKay, D.J.: Information-based objective functions for active data selection. Neural computation **4**(4), 590–604 (1992)
12. Merler, M., Galleguillos, C., Belongie, S.: Recognizing groceries in situ using in vitro training data. In: CVPR. IEEE Computer Society (2007)
13. Mikolajczyk, K., Schmid, C.: Scale & affine invariant interest point detectors. International journal of computer vision **60**(1), 63–86 (2004)
14. Singh, S., Gupta, A., Efros, A.: Unsupervised discovery of mid-level discriminative patches. Computer Vision–ECCV 2012 pp. 73–86 (2012)
15. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. pp. 3360–3367. IEEE (2010)
16. Yao, B., Khosla, A., Fei-Fei, L.: Combining randomization and discrimination for fine-grained image categorization. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. pp. 1577–1584. IEEE (2011)
17. Zhang, Y., Wang, L., Hartley, R., Li, H.: Where's the weet-bix? In: Asian Conference on Computer Vision. pp. 800–810. Springer (2007)