ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/352840003

Products-6K: A Large-Scale Groceries Product Recognition Dataset

Conference Paper · June 2021

DOI: 10.1145/3453892.3453894

CITATIONS
5

READS
277

11 authors, including:

Kostas Georgiadis
Information Technologies Institute (ITI)
30 PUBLICATIONS **345** CITATIONS
SEE PROFILE

Giorgos Kordopatis-Zilos
Czech Technical University in Prague
40 PUBLICATIONS **432** CITATIONS
SEE PROFILE

Fotis Kalaganis
Information Technologies Institute (ITI)
25 PUBLICATIONS **153** CITATIONS
SEE PROFILE

Panagiotis Migkotzidis
The Centre for Research and Technology, Hellas
9 PUBLICATIONS **34** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project H2020 - MAMEM - Multimedia Authoring and Management Using your Eyes and Mind View project

Project CUTLER - COASTAL URBAN DEVELOPMENT THROUGH THE LENSES OF RESILIENCY View project

All content following this page was uploaded by Giorgos Kordopatis-Zilos on 19 July 2022.

The user has requested enhancement of the downloaded file.

Products-6K: A Large-Scale Groceries Product Recognition Dataset

Kostas Georgiadis [*]
Information Technologies Institute,
Centre for Research and Technology
Hellas, Thermi 57001, Greece
kostas.georgiadis@iti.gr

Giorgos Kordopatis-Zilos
Information Technologies Institute,
Centre for Research and Technology
Hellas, Thermi 57001, Greece
georgekordopatis@iti.gr

Fotis P. Kalaganis
Information Technologies Institute,
Centre for Research and Technology
Hellas, Thermi 57001, Greece
fkalaganis@iti.gr

Panagiotis Migkotzidis
Information Technologies Institute,
Centre for Research and Technology
Hellas, Thermi 57001, Greece
migkotzidis@iti.gr

Elisavet Chatzilari
Information Technologies Institute,
Centre for Research and Technology
Hellas, Thermi 57001, Greece
ehatz@iti.gr

Valasia Panakidou
D. Masoutis S.A.
bpanikidou@masoutis.gr

Kyriakos Pantouvakis
D. Masoutis S.A.
kpantouvakis@masoutis.gr

Savvas Tortopidis
D. Masoutis S.A.
stortopidis@masoutis.gr

Symeon Papadopoulos
Information Technologies Institute,
Centre for Research and Technology
Hellas, Thermi 57001, Greece
papadop@iti.gr

Spiros Nikolopoulos
Information Technologies Institute,
Centre for Research and Technology
Hellas, Thermi 57001, Greece
nikolopo@iti.gr

Ioannis Kompatsiaris
Information Technologies Institute,
Centre for Research and Technology
Hellas, Thermi 57001, Greece
ikom@iti.gr

ABSTRACT

Product recognition is a task that receives continuous attention by the computer vision/deep learning community mainly with the scope of providing robust solutions for automatic checkout supermarkets. One of the main challenges is the lack of images that illustrate in realistic conditions a high number of products. Here the product recognition task is perceived slightly differently compared to the automatic checkout paradigm but the challenges encountered are the same. The setting under which this dataset is captured is with the aim to help individuals with visual impairment in doing their daily grocery in order to increase their autonomy. In particular, we propose a large-scale dataset utilized to tackle the product recognition problem in a supermarket environment. The dataset is characterized by (a) large scale in terms of unique products associated with one or more photos from different viewpoints, (b) rich textual descriptions linked to different levels of annotation and, (c) images acquired both in laboratory conditions and in a realistic supermarket scenario portrayed in various clutter and lighting conditions. A direct comparison with existing datasets of this category demonstrates the significantly higher number of the available unique products, as well as the richness of its annotation enabling different recognition scenarios. Finally, the dataset is also benchmarked using various approaches based both on visual and textual descriptors

CCS CONCEPTS

• **Human-centered computing**; • **Accessibility**; • **Accessibility systems and tools**;

KEYWORDS

Product Recognition, Groceries Dataset, Image Retrieval, OCR

ACM Reference Format:
Kostas Georgiadis, Giorgos Kordopatis-Zilos, Fotis P. Kalaganis, Panagiotis Migkotzidis, Elisavet Chatzilari, Valasia Panakidou, Kyriakos Pantouvakis, Savvas Tortopidis, Symeon Papadopoulos, Spiros Nikolopoulos, and Ioannis Kompatsiaris. 2021. Products-6K: A Large-Scale Groceries Product Recognition Dataset. In *The 14th PErvasive Technologies Related to Assistive Environments Conference (PETRA 2021), June 29–July 02, 2021, Corfu, Greece.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3453892.3453894

1 INTRODUCTION

According to the World Health Association (WHO), visual impairment is a major global health issue that affects more than 285 million individuals globally. The limitations imposed in terms of autonomy

[*] Corresponding Author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

PETRA 2021, June 29–July 02, 2021, Corfu, Greece
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8792-7/21/06…$15.00
https://doi.org/10.1145/3453892.3453894

in the everyday activities of the visually impaired, can be identified as one of the most crucial challenges associated with the specific health issue. One typical daily activity characterized by limited autonomy is the supermarket visit. The common practice for the visually impaired is to either visit the supermarket accompanied by a seeing person who will assist in the identification and purchase of goods or to provide their shopping list to a supermarket employee who will handle the product "retrieval". In the same context, Be My Eyes[1] connects visually impaired with sighted volunteers that provide visual assistance via live video calls, subject to the sighted users' availability. However, such practices significantly limit the autonomy of the visually impaired, as they rely on other individuals. Existing assistive devices either using mobile (e.g. Envision[2]) or glasses (e.g. OrCam MyEye 2[3]) based applications partly alleviate the problem, as the provided information is limited and mainly oriented to the task of object recognition (e.g. bottle, can), without offering specific information about identified products. Another alternative is Microsoft's Seeing AI[4] that scans the product's barcode and provides responses of high precision, nevertheless it would be challenging for the visually impaired to "locate" the barcode even when audio cues are provided. As a result, a system that will be able to adequately identify products would be an ideal assistive toolkit for the visually impaired for the supermarket scenario.

Creating such a system demands a detailed image dataset, capturing a large number of unique grocery products. Datasets in the supermarket context do exist, but are mainly oriented in tackling problems of different nature. The Freiburg Groceries Dataset [1] provides a total of approximately 5.000 images of 25 grocery classes, characterized by a large variety in terms of perspective, lighting and degree of clutter, aiming to tackle the problem of recognizing the object category in a real-life setting. The Grozi Dataset [2] includes a total of 120 unique products linked with images taken both in ideal conditions and in natural environments, with the task being the identification and localization of products of images in the latter category. The Supermarket Produce Dataset [3] is another grocery dataset that is built to facilitate the classification task of fruits and vegetables and consists of 2.633 images for 15 product categories. The RPC dataset [4] aims to tackle the problem of automatic checkout and provides an image database generated from 200 unique products. The images are captured in laboratory conditions with a wide variability in terms of perspective, resulting in 53.739 images of 200 isolated products and 30.000 images constructed for automatic checkout purposes. The SOIL-47 dataset [5] includes images of 47 products captured from various angles, with the scope of evaluating color-based recognition algorithms. The Grocery Products Dataset [6], another grocery dataset for product recognition, contains 8.350 candidate web-crawled frontal side images and 680 query images captured in natural shelf conditions, that correspond to 80 unique products.

The aforementioned datasets provide a high number of images (both target and query) that are usually captured under various view points, nevertheless the products identified by a unique Stock Keeping Unit (SKU) [7] are limited to a couple of hundreds (see Table 1). As a result embodying such datasets in a real-life scenario, where a system would have to be able to recognize products among thousands of unique SKUs, seems impractical. To this end and aiming at the increased independence of the visually impaired, we propose here the Products-6K dataset that is specifically designed to address the issue of product recognition in real-life scenarios, consisting of 6.348 unique SKUs. Each SKU is linked to candidate and query images captured under various conditions and viewpoints and is also associated with rich textual descriptors corresponding to different annotation levels provided by the supermarket's structured product categorization.

In this paper we propose a dataset, called hereby the Products-6K dataset, to support the task of large-scale product recognition. In essence, we consider the scenario where visually impaired individuals reach and grab a product from the shelf, aiming to identify whether the product at hand is the desired one. It is important to note here that we are based on the assumption that individuals have managed to approach a specific trail/shelf of interest using an assistive system as the one described in our previous work [8] and are about to select a specific product. The product recognition process is emulated here by capturing its image using the mobile phone's camera. Ideally, a product recognition system would provide the exact product's description with the consumer deciding if the selected product is the desired one. In cases where the system cannot provide a confident response about the exact product's description it can deliver more broad information like product's category or brand, that can be conveniently associated with the shelf/trail the individual is in front of.

2 DATASET CHARACTERISTICS

The Products-6K dataset is characterized by its large scale in terms of unique SKUs, as it provides 12.917 candidate images associated with 6.348 unique SKUs and 373 query images linked with 104 unique SKUs. The available images of both the candidate and the query set include single-product illustrations, with the former being captured in laboratory conditions and the latter being acquired in a supermarket environment. This partition was driven by the main goal of our dataset, which is to rely on an existing product's database that most supermarkets generate in laboratory conditions, so as to facilitate the product recognition task in a realistic supermarket scenario. Moreover, in an effort to mimic such a realistic scenario, the query images were acquired from several viewpoints under various conditions in terms of lighting and clutter, considering that the product has been removed from the shelf and its being held by the user in his/her hand. The different viewpoints and the various lighting/clutter levels are in essence the two major challenges for the efficient product recognition proposed dataset. Finally, each image (of both partitions) is accompanied by textual annotation in various levels of detail, including the product's SKU, its specific and broad product category and its brand that is stored in a .xls file.The dataset can be accessed via Zenodo[5] For reasons of corporate interest, the candidate and query images of 500 and 104 unique SKUs are available, but the interested reader is encouraged to follow the contact details for requesting access to the full dataset.

[1] https://www.bemyeyes.com/

[2] https://www.letsenvision.com/

[3] https://www.orcam.com/en/

[4] https://www.microsoft.com/en-us/ai/seeing-ai

[5] Products-6K: A Large-Scale Groceries Product Recognition Dataset - https://zenodo.org/record/4428917

Table 1: Dataset Comparisons

	Freiburg Groceries	Grozi	Supermarket Produce	RPC	SOIL-47	Grocery	Product–6K
# SKUs	25	120	15	200	47	80	6.348
# Images	5000	11.870	2.633	83.739	987	9.030	13.290

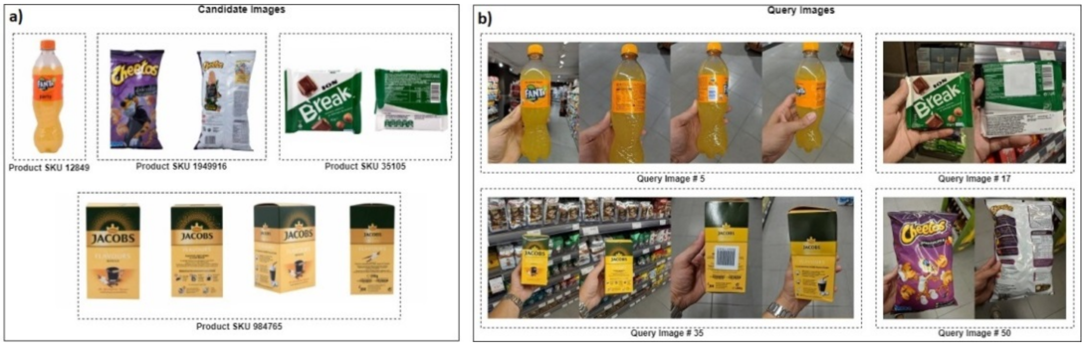

Figure 1: Sample images from (a) the set of candidate images and (b) the set of query images.

2.1 Construction of the candidate image set

The set of candidate images was provided by the Greek supermarket Masoutis[6], with a total of 12.917 product images being available, corresponding to 6.348 unique product SKUs. Each image associated with a specific SKU provided a different product view, with an example illustrated in Figure 1a. The frontal view was available for all SKUs, while the availability of other views (e.g. left/right side view) varied among SKUs. Each product image was linked with the product's SKU and accompanied by textual information corresponding to different levels of detail, starting from broad descriptions and ending up to a specific product description. The first annotation level provides information about the product's broad category (BPC), the second describes the product's specific category (SPC), the third determines its brand category (BC), while the last level presents the exact product description (PD). For example, for Product SKU 18283 (Fig. 1a), the product's BPC would be Coffee, its SPC Filtered Coffee, its BC Jacobs and its PD would be Coffee Jacobs Vanilla 250 grams.

To further enhance the available textual information we employed the optical character recognition (OCR) mechanism provided by the Google Cloud Vision API as a means to extract the accessible text from the product packages. The acquired textual information for each product was concatenated with the product's PD, resulting in an enhanced description of higher descriptive power (compared to PD), denoted as EPD. Consequently the EPD for Product SKU 18283, would be Coffee Filtered Jacobs Vanilla Flavours 250 grams.

2.2 Construction of the query image set

The query images were captured under real supermarket conditions using a mobile phone camera and mimicking the scenario where visitors grab a product from the shelf with their hand to decide whether or not to put it in their cart. Google Pixel 2 XL with a camera resolution of 12.2 MP (3024 x 4032 resolution) was selected for acquiring the query images. Images were taken under different real-world lighting conditions and degree of background clutter (see Figure 1b), aiming to emulate in the best way the scenario of product selection from the visually impaired. Considering that the (majority of) visually impaired will not be able to identify a specific view of the product package (e.g. the frontal view), the set of query images was enriched with extra views for each product. The number of the captured images or views for each product was dictated by the type of the product package. The physical limitations imposed by slim and bag packages resulted in capturing only 2 images for products arising from these categories, while for the remainder of package types (e.g. box, can, bottle) 4 images were registered. An illustration of four different package types, arising from the query images and corresponding to the same SKUs illustrated in Figure 1a, is available in Figure 1b. The visual inspection of Figure 1 readily reveals the differences between the images captured in laboratory (i.e. candidate images) and natural (i.e. query images) environment. The set of query images was created by identifying a subset of the candidate SKUs, consisting of 104 randomly selected unique SKUs, resulting in a total of 373 query images. Finally, the annotation detail level for each query image was identical to the one for the candidate image set (i.e. BPC, SPC, BC, PD and EPD).

3 EVALUATION PROTOCOL

In this section, a series of evaluation metrics for the task of product recognition are proposed. Prior to their description, it is important to note that only the set of query images must be used for evaluation purposes, considering that they have been obtained as a means to handle a specific problem (i.e. product recognition in realistic conditions). Additionally, considering that usually the frontal view of a product is much more informative, we also report accuracies for frontal-only views of the 104 products. Given the abundance of textual information being available for each SKU/image any

[6] https://www.masoutis.gr/

proposed product recognition system can be evaluated through a series of metrics formulated as follows:

SKU Accuracy (ACC_{SKU}): Measures the system's ability to reliably detect the exact product (i.e. SKU).

Broad Product Category Accuracy (ACC_{BPC}): Evaluates the system's ability to identify the BPC for the query image. Its impact is not as direct as in the case of ACC_{SKU}, but it can be equivalently assistive for the visually impaired providing more general information that can be associated with the trail the user is in front of.

Specific Product Category Accuracy (ACC_{SPC}): Metric similar to the ACC_{BPC} as it refers to the product's category but in a more detailed manner that is equivalent to the shelf that the user is in front of.

Brand Category Accuracy (ACC_{BC}): Evaluates the system's ability to determine the image's brand.

Finally, given that the task is an (image) retrieval problem, the aforementioned metrics reflect in essence the Precision at 1 (P@1) measurements. Aiming to also appraise the prospects of the provided dataset the cut-off ranks of 5/10/20, estimating the P@5, P@10 and P@20 respectively, have also been employed.

4 BENCHMARKING THE PROPOSED DATASET

This section demonstrates the use of the proposed dataset and also provides baselines for possible future comparisons. To benchmark the dataset, we employed both visual/textual descriptors and their combinations that are briefly described in the next sections. In terms of preprocessing applied to each query image independently, the first step was to locate the product in the provided image. The object covering the largest area must coincide, for the majority of the cases, with the product that must be retrieved. This is equivalent to performing object detection in the image and identifying the biggest bounding box (BB). Once the biggest BB was identified, by using its 2D coordinates to estimate the corresponding area, the second step consisted of using its coordinates to isolate it and crop it from the original image, with the cropped version being considered as the query image.

4.1 Visual Descriptors

Several relevant works have been proposed in the field of image retrieval. Early image retrieval works relied on solutions based on local handcrafted features, such as SIFT [9], in combination with some aggregation schemes, i.e., Bag-of-Wards [10], VLAD [11], or Fisher Vectors [12], in order to extract a vectorized representation for the images. More recent approaches adopt deep learning (DL) for feature extraction that significantly boost retrieval performance. Early methods that incorporated DL [13], [14] employed a pre-trained Convolutional Neural Network (CNN) and applied aggregation functions on its output. More recent methods fine-tune the CNN network on the particular problem they encounter, using sophisticated network architectures [15], [16] and loss functions [17]. Additionally, the state-of-art methods use global image representations disregarding the spatial dimension, which could be exploited for more accurate similarity calculations. To this end, in this work, we employ a solution based on visual features extracted from a pre-trained CNN and use symmetric Chamfer Similarity that considers spatial structure during the calculation [17].

More specifically, given an input image, a feature extraction scheme is applied to generate a region-level image representation. More precisely, features are extracted from the intermediate convolutional layers of a CNN [18], pre-trained on Imagenet [19][19], by applying Regional-Maximum Activations of Convolution (R-MAC) [14] on a specific granularity level. In that way, region-level representations that contain the spatial information of the images are generated. Unlike other feature extraction approaches, aggregation schemes are not applied on the spatial dimension of the extracted representations because this information is employed during similarity calculation. For the latter process, the image similarity using Chamfer Similarity [20] is calculated. Given a query and a target image, the similarity between every region pair of the two images based on their dot product is estimated, and then the similarity of the most similar regions in the target image for each region in the query image is averaged. In this work, the symmetric variant of Chamfer Similarity [20] is employed as it presents a good tradeoff between the preservation of image structure and invariance to spatial transformations (e.g., rotations, spatial shifts) that naturally exist in our particular problem setting. Due to space limitations, the reader is referred to [20] for a more detailed description about the similarity functions.

Based on the previously described schemes, the image indexing endpoint extracted features for candidate set, that were saved to an index along with the corresponding metadata information. The image search endpoint for near-duplicates then calculated the similarity between all index images with a given query image and ranked the results in descending order, with the top 20 images being selected for further processing.

4.2 Textual Descriptors

The majority of the product packages in our use case scenario contains significant amount of textual information. In this direction, besides the visual features derived using the near-duplicate detection (NDD) component, Google's OCR mechanism was employed to obtain the available text from each query and candidate image and was considered as an alternative descriptor in the product recognition process. A ranking process that included a distance metric between the OCR obtained text and the two different product descriptions (i.e. PDs and EPDs) that identified the common words in the two sets was then used to determine the SKU of each query image.

4.3 Fusion of visual and textual information

Additionally, approaches using both descriptors were explored. The fusion was performed in two stages and it boils down to a re-ranking process dependent to the second modality that will rerank the responses of the first modality and will provide a response regarding the query's SKU. In essence, when the initial ranking is provided by the NDD approach (i.e. first retrieval stage) the re-ranking (i.e. second retrieval stage) is performed by the OCR and vice versa. The two fusion approaches are referred hereafter as NDD+OCR and OCR+NDD respectively, with the first term denoting the preceding stage and the second the subsequent.

Table 2: The average accuracy scores (%) for the ACCSKU metric for all the baseline modalities.

	$\mathrm{OCR_{PD}}$	$\mathrm{OCR_{PD}}$+ NDD	$\mathrm{OCR_{EPD}}$	$\mathrm{OCR_{EPD}}$+ NDD	NDD	NDD+$\mathrm{OCR_{PD}}$	NDD+$\mathrm{OCR_{EPD}}$
All Images	18.18	30.91	39.27	32.90	34.55	37.88	41.69
Frontal Only	29.81	50.96	64.42	50.00	54.81	56.73	66.36

5 EXPERIMENTAL EVALUATION

The experimental evaluation section is divided in four subsections, with the first three incorporating the quantitative/qualitative experimental evaluation and the last describing the dataset's challenges.

5.1 Baseline Comparisons

This section reports the experimental results using the $\mathrm{ACC_{SKU}}$ metric as a means to evaluate the four different schemes (i.e. OCR+NDD, OCR, NDD and NDD+OCR) in terms of product (i.e. SKU) recognition. The evaluation process was performed independently for the full set of query images and for the frontal-only query images. As shown in Table 2, the sole use of the OCR modality using the product description (i.e. $\mathrm{OCR_{PD}}$) results in poor performance in terms of SKU retrieval, with the accuracy score barely surpassing 30% in the case of frontal query images, a fact that can be attributed to either the product's peculiar fonts or to the its decorative layout. Employing the NDD mechanism to re-rank the retrieved images provides a slightly improved performance, nevertheless, the obtained results are still inferior compared to both its enhanced version (i.e. $\mathrm{OCR_{EPD}}$) and the sole use of the visual descriptors (i.e. NDD) that reach approximately 39% and 35% for the full query image set. The approach of NDD+OCR is the scheme that performs better, with the accuracy for the full query set being 37.88% and 41.69% when the PD and EPD descriptors are employed. The enhanced product descriptions aided significantly in the performance improvement regarding the frontal query images, with the score for the NDD+$\mathrm{OCR_{EPD}}$ being 66.36% compared to the 56.73% of the NDD+$\mathrm{OCR_{PD}}$. A noticeable fact here is the significantly better performance for the retrieval of frontal images regardless of the selected modality, that can be attribute to a series of factors discussed in Section 5.4.

5.2 Annotation Detail Comparisons

Given the established superiority of fused modality of NDD+$\mathrm{OCR_{EPD}}$ the remainder of the accuracy metrics is examined only for this modality and is tabulated in Table 3. The scores obtained for the products' categories and brand are significantly higher compared to the $\mathrm{ACC_{SKU}}$ an outcome that can be attributed to the high number of products belonging to the same product family, concerning either its category or brand. It is noteworthy, that the P@1 scores for the frontal-only query images reach near optimal responses with the products' categories or brand being correctly retrieved approximately nine out of ten times. The high scores obtained for P@5, P@10 indicate the large room for improvement for the proposed dataset especially regarding the query version that considers all images. Moreover, the difference observed between P@1-P@5 and P@5-P@10 suggests that the requested information usually lies within the first 5 or 10 retrievals and that alternative re-ranking approaches could significantly increase the P@1 scores, while the marginal difference between P@10 and P@20 pinpoints that retrieving the top 10 images would suffice for most cases.

5.3 Qualitative Evaluation

This section provides indicative examples of the retrieval results of NDD+$\mathrm{OCR_{EPD}}$. Figure 2, illustrates correctly retrieved images in the first rank for various types of packages, while Figure 3 cases where the correct candidate image lies within the first 5 ranks. It is evident that for the majority of the examples, the retrieved images are similar, making the selection of the correct one a challenging task. Additionally, the observed tendencies can justify the increased scores observed when comparing P@1 with P@5.

5.4 Challenges encountered

During the creation and evaluation process of the proposed dataset, various challenges associated both with the dataset itself and the marketing area in general have been encountered. As described in previous sections, there were cases (i.e. SKUs) in the set of target images that were linked only with frontal images or with a limited number of views. This is expected to have a direct impact on the NDD-based retrieval process, as such query images will not be retrievable. A typical example of the specific challenge is depicted in Figure 4a, where for a specific SKU both its target and query images are presented in the left and right panel respectively, showcasing the aforementioned challenge. It is rather unlikely to have side-views of this kind in the product dataset of a supermarket, but they are expected to appear as query images given the assumption that visually impaired are not always able to determine the desired view-point.

In the same direction, SKUs that differ only in the frontal view of the product package were encountered. In this case, products belonged to the same brand and possibly product category and their side views were identical resulting in various misclassified images. Two such cases are available in Figure 4b, where two query images associated with different SKUs are almost identical in terms of context. The issues posed by these two challenges can be readily depicted in the substantial gap in terms of accuracy score, encountered between the full query set and the frontal-only images.

The last challenge is attributed to the ever evolving nature of the marketing area that dictates frequent alteration in the product packages aiming in enhanced attractiveness and increased sales. As a consequence, there were a few cases in the dataset where the available candidate images were different than the query images due to package alteration. Cases with minor package alterations did not seem to affect the NDD retrieval process, nevertheless when changes in the package were major the retrieval process in several cases. Figure 4c and Figure 4d illustrate examples of minor changes that were correctly retrieved and of major changes between the two

Table 3: Precision at 1/5/10/20 for the four variants of the accuracy.

	All Images				Frontal Only			
ACC_{SKU}	P@1	P@5	P@10	P@20	P@1	P@5	P@10	P@20
ACC_{BPC}	41.69	53.78	62.24	66.36	66.36	81.73	85.58	85.58
ACC_{SPC}	63.94	70.61	75.15	79.09	89.42	92.31	92.31	92.31
ACC_{BC}	58.18	70.00	74.55	79.09	88.46	93.27	93.27	93.27
ACC_{SKU}	64.85	76.67	80.61	86.67	91.35	96.15	96.15	97.12

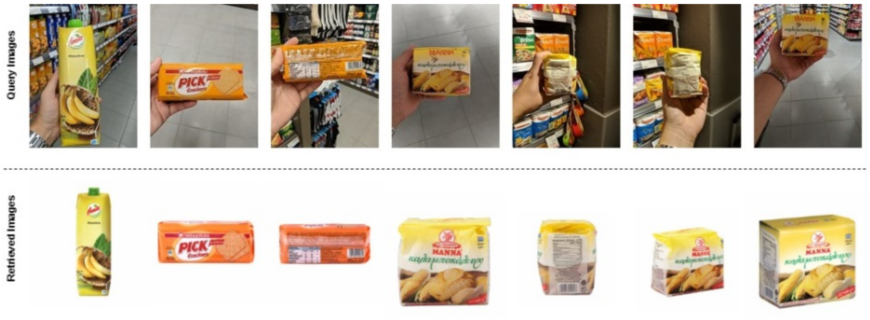

Figure 2: Exemplar cases of correctly retrieved query images.

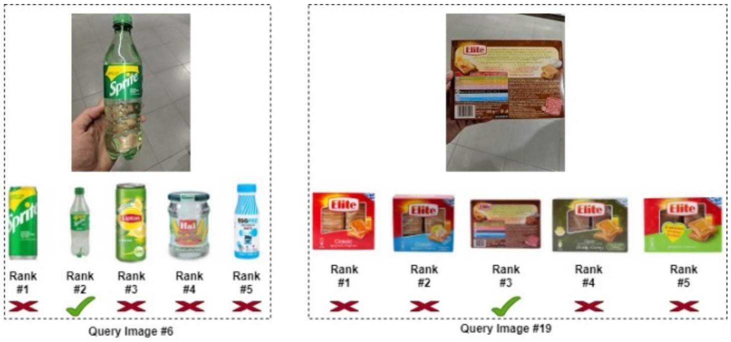

Figure 3: Exemplar cases where the query image is identified within the five first ranks of the retrieved images.

images (i.e. target and query) that results in the misidentification of the latter respectively. This challenge is expected to self-regulate in a real world application, considering that supermarkets update their product database on a regular basis with images depicting the new packages.

6 DISCUSSION AND CONCLUSIONS

In this paper the Products-6K dataset was introduced, a novel dataset for large scale product recognition in a supermarket environment. The dataset encompasses a total of 6348 unique SKUs and approximately 12.917 candidate product images and 104 SKUs-373 product query images for evaluation purposes. Besides the product description for each image/SKU the categorization in terms of broad, specific and brand category is also provided. The candidate images were captured in laboratory conditions, while the query images were captured in real supermarket conditions under various viewpoints and degrees of background clutter/lighting conditions. The dataset aims to tackle the problem of large scale product recognition as encountered by the visually impaired. The problem is challenging, considering that the number of the available SKUs is significantly higher compared to previously released datasets. The challenging nature of this dataset can be readily recognized by the accuracy scores obtained by the benchmarking approaches in the product recognition task (i.e. ACC_{SKU}), with scores barely surpassing 40% and 65% for the full query set and for the frontal-only images respectively. The accuracy scores obtained for the category/brand recognition task (i.e. ACC_{BPC}, ACC_{SPC} and AC_{CBC}) were significantly higher, with frontal-only images being correctly associated with the corresponding category/brand with 90% accuracy. The difference observed between the scores of P@1 and P@5 indicate the room for substantial improvement in the proposed dataset and that it can be employed in several other future research directions, including among others automatic checkout and brand detection. Additionally, the challenges of this dataset

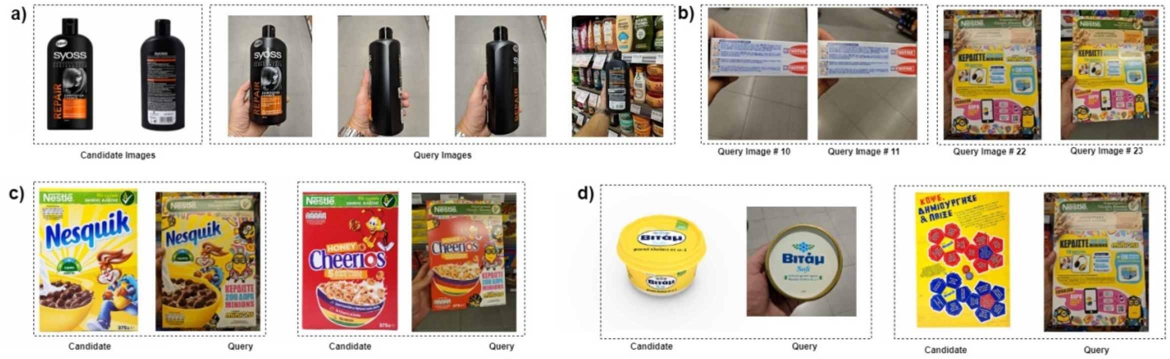

Figure 4: Examples of Dataset Challenges. a) SKU with limited viewpoints. b) Identical package illustrations associated with different query images and thus SKUs. c) Minor Package Alterations. d) Major Package Alterations.

must be taken under consideration that mainly derive from the discrepancies between the candidate and query image set, with the former being generated in laboratory conditions mainly to support the supermarket's e-shop, while the latter were acquired in realistic shopping scenarios. Finally, an issue left untreated in the evaluation of the proposed dataset is its evaluation in different crop degrees, considering that the visually impaired may not be able to capture the entire package.

ACKNOWLEDGMENTS

This work is part of project e-vision that has been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (code:T1EDK-02454).

REFERENCES

[1] Jund, P., Abdo, N., Eitel, A., & Burgard, W. (2016). The freiburg groceries dataset. arXiv preprint arXiv:1611.05799.
[2] Merler, M., Galleguillos, C., & Belongie, S. (2007, June). Recognizing groceries in situ using in vitro training data. In 2007 IEEE Conference on Computer Vision and Pattern Recognition (pp. 1-8). IEEE.
[3] Rocha, A., Hauagge, D. C., Wainer, J., & Goldenstein, S. (2010). Automatic fruit and vegetable classification from images. Computers and Electronics in Agriculture, 70(1), 96-104.
[4] Wei, X. S., Cui, Q., Yang, L., Wang, P., & Liu, L. (2019). RPC: A large-scale retail product checkout dataset. arXiv preprint:1901.07249.
[5] Koubaroulis, D., Matas, J., Kittler, J., & CMP, C. (2002, January). Evaluating colour-based object recognition algorithms using the soil-47 database. In Asian Conference on Computer Vision (Vol. 2).
[6] George, M., & Floerkemeier, C. (2014, September). Recognizing products: A per-exemplar multi-label image classification approach. In European Conference on Computer Vision (pp. 440-455). Springer, Cham.
[7] Van Kampen, T. J., Akkerman, R., & van Donk, D. P. (2012). SKU classification: a literature review and conceptual framework. International Journal of Operations & Production Management.
[8] Georgiadis, K., Kalaganis, F., Migkotzidis, P., Chatzilari, E., Nikolopoulos, S., & Kompatsiaris, I. (2019). A Computer Vision System Supporting Blind People-The Supermarket Case. In International Conference on Computer Vision Systems (pp. 305-315). Springer.
[9] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. International journal of computer vision, 60(2), 91-110.
[10] Sivic, J., & Zisserman, A. (2003). Video Google: A text retrieval approach to object matching in videos. In null (p. 1470)..
[11] Jégou, H., Douze, M., Schmid, C., & Pérez, P. (2010). Aggregating local descriptors into a compact image representation. In 2010 IEEE computer society conference on computer vision and pattern recognition (pp. 3304-3311).
[12] Perronnin, F., Liu, Y., Sánchez, J., & Poirier, H. (2010). Large-scale image retrieval with compressed fisher vectors. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (pp. 3384-3391).
[13] Babenko, A., & Lempitsky, V. (2015). Aggregating local deep features for image retrieval. In Proceedings of the IEEE international conference on computer vision (pp. 1269-1277).
[14] Tolias, G., Sicre, R., & Jégou, H. (2016). Particular object retrieval with integral max-pooling of CNN activations. In International Conference on Learning Representations.
[15] Radenović, F., Tolias, G., & Chum, O. (2018). Fine-tuning CNN image retrieval with no human annotation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(7), 1655-1668.
[16] Gordo, A., Almazan, J., Revaud, J., & Larlus, D. (2017). End-to-end learning of deep visual representations for image retrieval. International Journal of Computer Vision, 124(2), 237-254.
[17] Revaud, J., Almazán, J., Rezende, R. S., & Souza, C. R. D. (2019). Learning with average precision: Training image retrieval with a listwise loss. In Proceedings of the IEEE International Conference on Computer Vision (pp. 5107-5116).
[18] Kordopatis-Zilos, G., Papadopoulos, S., Patras, I., & Kompatsiaris, Y. (2017). Near-duplicate video retrieval by aggregating intermediate CNN layers. In International conference on multimedia modeling (pp. 251-263).
[19] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.
[20] Kordopatis-Zilos, G., Papadopoulos, S., Patras, I., & Kompatsiaris, I. (2019). ViSiL: Fine-grained spatio-temporal video similarity learning. In Proceedings of the IEEE International Conference on Computer Vision (pp. 6351-6360).

View publication stats