A DATASET FOR HAND-HELD OBJECT RECOGNITION

Jose Rivera-Rubio *Saad Idrees* *Ioannis Alexiou* *Lucas Hadjilucas* *Anil A. Bharath*

Imperial College London

ABSTRACT

Visual object recognition is just one of the many applications of camera-equipped smartphones. The ability to recognise objects through photos taken with wearable and handheld cameras is already possible through some of the larger internet search providers; yet, there is little rigorous analysis of the quality of search results, particularly where there is great disparity in image quality. This has motivated us to develop the Small Hand-held Object Recognition Test (SHORT). This includes a dataset that is suitable for recognising hand-held objects from either snapshots or videos acquired using hand-held or wearable cameras. SHORT provides a collection of images and ground truth that help evaluate the different factors that affect recognition performance. At its present state, the dataset is comprised of a set of high quality training images and a large set of nearly 135,000 smartphone-captured test images of 30 grocery products. In this paper, we will discuss some open challenges in the visual object recognition of objects that are being held by users. We evaluate the performance of a number of popular object recognition algorithms, with differing levels of complexity, when tested against SHORT.

Index Terms— Dataset, object recognition, content-based image retrieval, assistive devices.

1. INTRODUCTION

In this paper, we present SHORT-30, an updated and practical dataset for studying object recognition and retrieval in the challenging scenarios of hand-held objects and mobile or wearable cameras, and with emphasis in the assistive case for blind and partially sighted users. It is open to the research community for download at [1] and the website includes collaborative tools to help expand SHORT. We also provide baseline performance measurements on the current dataset using object recognition algorithms of different degrees of recognition complexity.

One of the motivations for introducing SHORT is the possibility of applying visual object recognition to assistive applications, supporting people with visual impairment. The ubiquity of camera-equipped smartphones, and their accessibility options, make them potentially well suited to this application.

Image-based object recognition presents very particular challenges for assistive use, as the variability of viewing conditions (lighting, point of view, etc.) is large. Barcodes are not always easily located or read by users, and may also be vendor-specific. In addition, the precision of search, and guaranteed matches against curated, high-quality product information is essential: internet-trawled product images are not sufficiently controlled.

(a) Collage representing the grocery products in the SHORT-30 dataset. This is a selection of items to include cans (shiny), boxes, uneven surfaces, similar shapes, semi-transparent or deformable packaging. These are popular products that are widely available for easy reproducibility and contain snacks, toiletries, medicines, drinks, canned food, dairy products, etc.

(b) Sample test images. **Top row**: still-images; **bottom row**: video frames. Note: the images were cropped to fit the collage, they actually have different resolutions. This query selection contains samples from all the test datasets (see Table 1.)

Fig. 1. Sample of SHORT-30 training set (a) and test set (b).

2. SHORT-30 AND RELATED DATASETS

A related database of house-hold products is the Grozi-120 dataset [2]. It contains 120 categories of groceries and is divided into "model" and "query" sets. For every product, the models were downloaded from the Internet while the queries consisted of cropped video frames from recordings of supermarket shelves. SHORT provides a curated database of models to train object recognition algorithms taking also into account the assistive usage context, where the queries are highly variable. Grozi-120, however, lacks the variability and background clutter that would occur in real world scenarios, and the training images only have a limited number of views per product, usually just the frontal one showing the brand. SHORT expands the

number of views to 36, with 12 different levels of rotation and 3 elevations. The multiple views allow us to determine the importance of viewpoint in being able to guarantee a definitive match.

Caltech-101 and 256 [3, 4], together with PASCAL VOC [5] have been widely used to train and benchmark object recognition and detection algorithms. Caltech's dataset increased the depth of previous datasets, achieving a minimum of 80 images per category to favour the variability in training and test database size. However, neither dataset is recommended for localisation tests as the images contain "photographer's bias" in which the objects are usually placed near the center of the image. Nevertheless, the challenging nature of the PASCAL datasets, and the well-defined evaluation protocol, has led to it being a widely-cited benchmark for object recognition algorithms over recent years. However, only 2 out of 20 categories have a depth larger than 1,000 images per category, while in SHORT the minimum number of images per category is 3,507, outnumbering the latest PASCAL and both Caltech datasets.

ImageNet [6] was publicly released in 2010 to increase the number of categories up to the level of human recognition, which is estimated to be in the range of the tens of thousands [7]. ImageNet focuses on object categorisation "at near human scale" and therefore provides 1.2 million images of a broad range of objects belonging to 1,000 categories [3]. This remarkable dataset depth, however, presents certain disadvantages when the scope of the application is more specific, as it is in the context of shopping or assistive systems. The 30 products from SHORT-30, and the ones that will be obtained for subsequent expansions, are widely available. However, only 6 out of the 30 can be found in ImageNet: Coca-Cola, orange Fanta, semi-skimmed milk, orange marmalade, deodorant and OXO chicken cubes. From these some presented ambiguities: the category milk contained 231 images; some represented milk bottles valid for a shopping context, but some of them depicted a glass or jug of milk, milk crates in a factory, or other packages of milk. ImageNet cannot be used to train a system that guarantees a minimum scalability in a shopping context. Another criticism lies in the fact that some specific items are hard to index. In ImageNet, for instance, the orange Fanta is under drinks $\rightarrow$ soft drinks $\rightarrow$ orange synset. This makes it difficult to use ImageNet as a benchmark dataset for such a specific application.

Another important limitation of existing datasets is that they use large amounts of training data containing unsystematic views of an object to train classifiers; this introduces bias and can lead to "solving" the dataset. SHORT, however, offers a training set of model images which systematically cover variations in an object's viewing angle. This allows to study how recognition of queries taken non-systematically is affected by the variability in viewpoints in the training data.

Test images from many datasets lack the variability in views or introduce the photographer's bias. On the other hand, the image queries in SHORT have been captured by multiple users with a variety of the latest smartphone cameras covering a wide range of viewing angles and containing images at current typical resolutions. This context is not covered by traditional datasets, in which high quality catalogue images are being compared with variable quality user-captured images; this makes the matching more challenging in SHORT than other datasets. Images of similar quality are often not present in both "database" and "query" datasets, a situation being increasingly encountered in commercial applications.

As an additional feature SHORT also contains test images acquired by blindfolded users and therefore mimics scenarios involving visually impaired users.

3. SHORT-30 TECHNICAL DETAILS

3.1. Overview

SHORT is comprised of separate datasets for training and testing. Currently, the training dataset consists of high resolution acquisitions of 30 grocery items acquired in a very controlled setup, with 36 images of the same object from different angles and views. For testing, we provide a set of query images of the same grocery items acquired with 30 different smartphones. Lighting, pose, sensors and optics were quite varied. This represents a more realistic view of hand-held object queries from hand-held devices. The SHORT-30 dataset contains an average of more than 4,200 queries per product, allowing a realistic study of factors that affect recognition quality. In addition, video sequences of hand-held objects contain blur and different background clutter, as the volunteers moved while capturing sequences, relevant to a use case that might be considered as streamed object recognition.

In addition to the images, we provide ground truth annotations for all the data in terms of its object class label. We also include binary masks of the objects from the training dataset, indicating the bounding box around each item.

SHORT is openly available and it can be downloaded from [1] and the website is also a platform where database users and SHORT curators can interact. The first expansion of SHORT is taking place in June 2014 and takes into account feedback from the community on the usability of SHORT. The release also includes code and evaluation data.

3.2. Image acquisition protocol

3.2.1. Training images

The database of models was acquired with a Nikon D7000 SLR camera using a 18-105 mm lens connected to a laptop and using the Nikon *live capture* software. The 16.2 megapixel captures in *raw* format were kept, but a JPEG copy of each image was also generated with a resolution of 4928×3264 pixels. A 986×653 resized copy of the high resolution images is also made available

A total of 36 views were acquired per category. The views used in the product models contain shots at three elevations (17, 47 and 68 cm, at a distance of 1m) above the object base. 12 degrees of rotation were used per elevation. Both the background and a turntable containing the object were covered with a uniform "chroma blue" cloth, . A quartet of 11 watt PL fluorescent lamps were used to illuminate background and object. The process described above produces high-quality database (training) images; however we took no such precautions with the query (test) images. The difference in capture quality makes SHORT very relevant to the usage case that we have outlined, in which high-quality product images are used to provide a controlled database of items. We see this factor as very important in order to quarantee the quality of information. Such multiview images are routinely acquired for product catalogues, marketing brochures, and websites, forming a standard part of the product manufacturing processes.

3.2.2. Test images

Two experimental sessions were conducted. Volunteers were asked to take a minimum of five shots of every product and a five second video. No other instructions were given on how to acquire the images. During the second acquisition experiment, the images were taken with blindfolded users, reducing the alignment bias that

Dataset	Total Images	Images Per Category		
		Min	Max	Mean
ST-SG	2,797	75	115	93.23
VF-SG	91,293	2,115	4,033	3,043.10
ST-BF	1,225	32	57	40.83
VF-BF	39,209	832	1,884	1,306.97
All	134,524	3,054	6,089	4,213.9

Table 1. Summary of SHORT test datasets. Still images (ST) and videoframes (VF) acquired by sighted users (SG) or blindfolded (BF).

a sighted user might have; we used this as a proxy for the assistive device context. Visually-impaired users will also be recruited once the quality of search can be demonstrated as being sufficient for real-world use.

A total of 30 different camera-equipped smartphones was used, with resolutions ranging from 320×240 to 3264×2448 pixels. The variability of camera characteristics, parameters and capture conditions is enormous and a very distinctive feature of this dataset. The collage shown in Fig. 1b contains a small sample of the variability present in the test dataset. As can be appreciated, images contain different views, levels of sharpness, background clutter, occlusion, illumination, and specular reflection. These realistic features of the queries, together with the availability of sighted and blindfolded sets, help in identifying certain characteristics required for database quality and coverage in order to meet the assistive and hand-held usage contexts.

4. BENCHMARKS

4.1. Classification using SIFT Descriptor Matching

Object recognition systems can employ several properties of objects, ranging from colour distributions to texture and gradient field fingerprints, such as histograms of gradients (HoG). However, without employing more computationally expensive geometric validation of putative matches between object models and queries, visual word performance is likely to be at least partly determined by the discriminative power of the raw descriptors themselves. We therefore include a variant on pairwise SIFT [8] descriptor matching for generating scores between a query and each model image in the training dataset. Each match that passes Lowe's uniqueness criterion [8] is ranked according to a distance score μ; the match with the lowest distance score is assigned the highest ranked similarity. A query is then classified as belonging to the class $C^{(p)}$ corresponding to the training image $\mathbf{T}^{(p)}$ if

$$p = \arg\min_{n} \{\mu^{(n)}\} \tag{1}$$

with p being one of the indices of the training images $[1, ..., N]$.

This descriptor-based method is likely to be indicative of the discriminating capacity of individual descriptor types, having an effect on the ultimate performance of any object recognition technique. However, because it relies on descriptor-by-descriptor comparison, it is not readily scalable to large database sizes. Nevertheless, we keep it as one type of "gold standard" method and compare it with more scalable techniques in the following section.

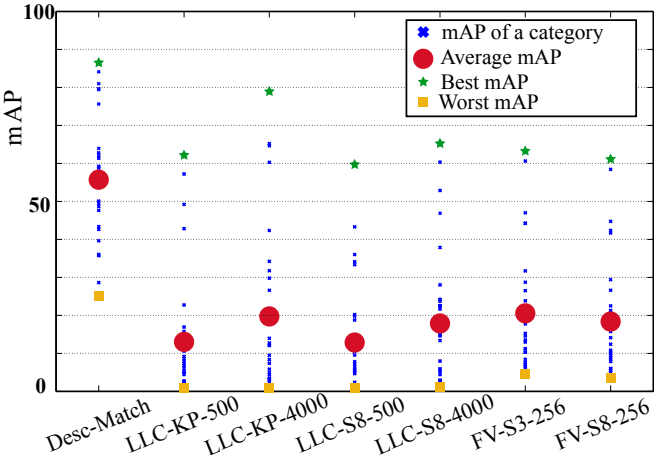

Fig. 2. Detailed performance evaluation of state-of-the-art algorithms on SHORT-30. **LLC** – locality-constrained linear coding; **FV** – Fisher vector. The SIFT descriptors can be computed on dense grids with a spacing of **Sx** pixels or around the SIFT keypoints (**KP**). The last figure indicates the size of the visual vocabulary (256, 500 or 4000 visual words).

4.2. Benchmarking: More Scalable Approaches

The challenges posed by this dataset was also assessed using a fairly standard recognition "pipeline" to provide category (product ID) ranking. The SIFT descriptor was applied in one of two approaches: dense or sparse. The dense approach employs a grid with either 3 or 8 pixel spacing, and a 16×16 spatial extent for the single-scale approaches. The sparse approach uses keypoint detection with standard SIFT-based scale-selection. The VLFEAT implementation is used for the descriptor-keypoint acquisition. Three different histogram encoding methods were applied: hard assignment [9] (using 400 and 4000 visual words), Locality-Constrained Linear coding (LLC) [10] and Fisher Vector (FV) encoding [11]. We used the setup of Chatfield et al. for the FV approach as it is known to perform best [12].

On top of LLC and FV, spatial pooling was applied, using the pyramid approach described in [13] and with 3 pyramidal levels (0,1,2). Kernels (as defined in [14]) were first computed for each pyramidal level [15]. Kernels were then fused and fed to an SVM to determine the classification accuracy and average precision.

5. RESULTS AND DISCUSSION

In order to facilitate performance comparisons, we organised SHORT queries into the four groups in Table 1. From Table 2 we appreciate that these groups experienced differences in average quality of match. Still image queries outperform single-frame queries taken from unfiltered video: video frames typically include images which are blurred due to an object being rotated during capture. However, pilot work suggests that multiple frames from video can enhance accuracy. Queries from sighted users led to better performance than those from blindfolded users. The "blindfolded" queries appeared less susceptible to some forms of capture-bias: some images contained inadvertent partial object occlusions, making categorisation more challenging. This makes SHORT arguably a better dataset for designing technology for vision-based assistive

systems for hand-held /wearable camera recognition of hand-held objects.

mAP	ST-SG	VF-SG	ST-BF	VF-BF
HA-4000	45.86	36.86	27.56	26.03

Table 2. Recognition performance across the four different subgroups of SHORT-30. The reduction in quality of match for the blindfolded subgroup is notable.

Classification results for the different methods are summarised in Fig. 2. The low performance of most state-of-the-art methods demonstrates the challenge presented by SHORT-30. The precision/recall analysis shown in Fig. 3 illustrates a remarkable variability in retrieval performance across categories. This fact reflects the complexity of the recognition problem, and the need for more robust algorithms.

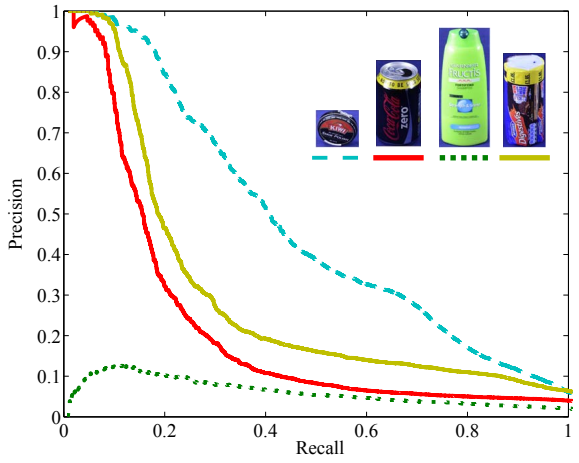

Fig. 3. LLC-S8-4000 Test. Representative empirical precision and recall curve for a small sample of product classes. Only four classes, including best and worst results, are represented to help visualisation. The test was run with 59,226 queries against the database of 1,080 models. Performance in all categories is summarised in Fig. 3

Recently, the value of well-defined, robustly labelled datasets in both evaluating and training object recognition systems has become clear. Table 3, provides a comparison of different datasets, and methods of recognition. A few interesting observations may be made. For example, HA-4000 yields high average precision in the SHORT dataset, but appears to yield lower performance in Caltech-101 and PASCAL VOC databases. However, the results are the other way around for HA-500. This suggests that a larger vocabulary is needed for the objects in SHORT. However, in the case of Caltech-101 where the images are of much lower resolution than SHORT, the details are not easily resolvable and hence a larger vocabulary captures irrelevant variations such as noise. Grozi-120 has a lower mAP than most of the SHORT groups, which is possibly due to its rather low spatial resolution. These observations suggest that we are still some way from having a single dataset that adequately represents all use cases for object recognition.

mAP	VOC-2007	Caltech-101	SHORT-30-ST	SHORT-30-VF
HA-4000	37.50	18.91	45.86	36.86
HA-500	-	61.20	24.54	22.39
LLC-S8-4000	46.01	66.64	17.89	11.20
FV-S8-256	59.35	77.78	18.39	-

Table 3. Dataset comparison. Classification results of baseline performance algorithms on SHORT-30 and other existing datasets.

6. CONCLUSION AND FUTURE WORK

We have presented a new publicly available dataset containing queries consisting of single images and video frames of hand-held objects. These were captured by multiple users using a variety of smartphones. The purpose of this dataset set is to have a more realistic view of variability in acquisition across images taken of hand-held objects. Unlike other datasets, SHORT also contains a great disparity in the image resolution and quality of its object and query items, which we feel is more representative of conditions of use we would expect for high-quality, curated models being used to service queries from wearable of hand-held cameras.

The training set of model images systematically captures variations in objects' viewing angle allowing us to study the effect of the number of viewing angles present in a database on matching quality in a widely varying query. SHORT can be used to develop object recognition algorithms targeted for both sighted and visually-impaired users: comparisons of such will allow a better understanding of the extra technical requirements placed on computer vision systems by users who may not be aware of the quality of the images they are capturing. Additionally, SHORT paves the way for addressing several open questions; for example:

a) How to minimise – or better yet, eliminate – the chance of a false positive match. Such errors could be dangerous when used in the context of household product identification. SHORT-30, with a range of variability in the test images, can be used to establish image quality metrics for accepting a query in this context.

b) How to ensure that ambiguity between two similar products might be noted and used to either eliminate a candidate search result as being too uncertain a match, or to prompt a user to submit specific queries based around competing hypotheses of what the object might be.

SHORT-30 was developed with the intention of piloting a larger dataset to test such strategies, and indeed our future work will aim to increase the number of classes. We are currently exploring curated crowdsourcing techniques to involve a larger community in growth of the dataset.

Our first step towards this goal is to involve the potential dataset users in its expansion. The download website [1] provides facilities for feedback. A second acquisition stage is planned, using responses from the research community to expand the current version of SHORT in order to address challenges either in scaling or in the details of usage context.

The shopping and assistive contexts pose a real challenge for the current generation of recognition algorithms, and SHORT-30 represents a key step towards being able to assess performance under realistic query conditions.

7. REFERENCES

[1] Jose Rivera-Rubio, Saad Idrees, and Anil A. Bharath, "SHORT dataset website," http://short.bicv.org.

[2] Michele Merler, Carolina Galleguillos, and Serge Belongie, "Recognizing groceries in situ using in vitro training data," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2007, pp. 1–8, IEEE.

[3] L Feifei, R Fergus, and P Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," *Computer Vision and Image Understanding*, vol. 106, no. 1, pp. 59–70, 2007.

[4] G Griffin, A Holub, and P Perona, "Caltech-256 object category dataset," Tech. Rep. 1, California Institute of Technology, 2007.

[5] Mark Everingham, Luc Gool, Christopher K I Williams, John Winn, and Andrew Zisserman, "The PASCAL Visual Object Classes (VOC) Challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2009.

[6] Jia Deng, Wei Dong, Richard Socher, and LJ Li, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 2–9.

[7] I Biederman, "Recognition-by-components: a theory of human image understanding.," *Psychological Review*, vol. 94, no. 2, pp. 115–47, Apr. 1987.

[8] David G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[9] Gabriella Csurka and C Dance, "Visual categorization with bags of keypoints," in *Workshop on Statistical Learning in Computer Vision, ECCV*, 2004, vol. 1, pp. 1–22.

[10] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong, "Locality-constrained Linear Coding for image classification," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. June 2010, pp. 3360–3367, IEEE.

[11] Florent Perronnin, J Sánchez, and Thomas Mensink, "Improving the fisher kernel for large-scale image classification," in *European Conference on Computer Vision (ECCV)*, 2010, pp. 143–156.

[12] Ken Chatfield, Victor Lempitsky, Andrea Vedaldi, and Andrew Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods," *Procedings of the British Machine Vision Conference 2011*, , no. 1, pp. 76.1–76.12, 2011.

[13] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer Vision and Pattern Recognition*. 2006, vol. 2, pp. 2169–2178, IEEE.

[14] A. Vedaldi and A. Zisserman, "Efficient additive kernels via explicit feature maps," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[15] Koen E A Van De Sande, Theo Gevers, and Cees G M Snoek, "Evaluating color descriptors for object and scene recognition.," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1582–1596, 2010.

[16] Jose Rivera-Rubio, Saad Idrees, Ioannis Alexiou, Lucas Hadjilucas, and Anil A. Bharath, "Mobile visual assistive apps: Benchmarks of vision algorithm performance," in *ICIAP 2013*, vol. 8158 of *Lecture Notes in Computer Science*, pp. 30–40. 2013.