PharmaPack: mobile fine-grained recognition of pharma packages

O. Taran, S. Rezaeifar, O. Dabrowski, J. Schlechten, T. Holotyak, S. Voloshynovskiy*

***Abstract*— We consider the problem of fine-grained physical object recognition and introduce a dataset PharmaPack containing 1000 unique pharma packages enrolled in a controlled environment using consumer mobile phones as well as several recognition sets representing various scenarios. For performance evaluation, we extract two types of recently proposed local feature descriptors and aggregate them using popular tools. All enrolled raw and pre-processed images, extracted and aggregated descriptors are made public to promote reproducible research. To evaluate the baseline performance, we compare the methods based on aggregation of local descriptors with methods based on geometrical matching.**

I. INTRODUCTION

Many multimedia and security applications require accurate recognition of physical objects using mobile phones of end users. These applications include *mobile shopping and visual search*, *objects tracking and tracing* including delivery and distribution chain control, generation of usage statistics, etc. and *anti-counterfeiting*. The latter includes the detection of fake objects to prevent their consumption and illegal distribution.

Pharmaceutical products, often distributed in packages, are very important groups of products for the following reasons. Counterfeit pharma products might contain dangerous components or lack the proper active ingredients. At the same time, being quite expensive they represent an attractive target for counterfeiters. In many cases, the consumers rely on the information printed on the packages and make their decision about the products authenticity considering the quality and presence of protection features on the packages. However, nowadays, the quality of reproduction techniques is extremely high and relatively cheap the fakes might be very close to the original ones. Moreover, it is very rare that an end consumer knows all details of the used protection to distinguish a fake without special training or special technical means. Therefore, the protection of pharma packages is a very important economic and social problem. The reliable recognition of physical object is the first step towards the protection of pharma packages and the creation of attractive mechanisms of interaction between the packages and end consumers. In turn, it also enhances the efficiency of their correct usage while at the same time leading to global tracking and tracing methods whilst hindering the world wide counterfeiting cartels.

This research was partially supported by the SNF project 200021E-164334.

* S. Voloshynovskiy is a contact author. O. Taran, S. Rezaeifar have the equal contribution. O. Dabrowski and J. Schlechten participated as part of their Master projects.

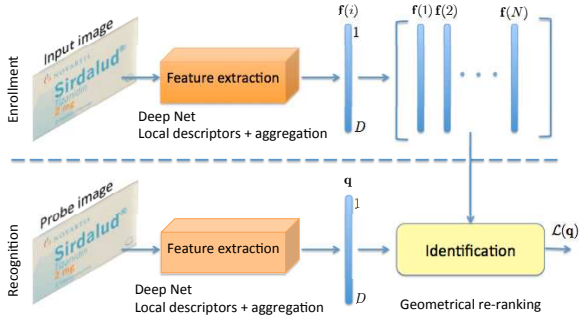


Fig. 1: Generalized diagram of pharma package recognition based on high dimensional features of dimension D extracted from images.

A. State-of-the-art in the package databases

It should be pointed out that the computer vision and pattern recognition community have developed many datasets targeting physical object recognition. Without pretending to be exhaustive in our overview, we mention that ALOI dataset containing 1000 objects [1], the ImageNet dataset containing 200 categories of objects [2], the dataset from the PASCAL 2012 challenge containing 20 classes of objects [3] and probably the closest to our application, the Stanford Mobile visual search data set contains 23 different objects such as books, CD covers, DVD covers and common objects [4].

However, up to our best knowledge, there does not exist any public database containing a sufficient number of objects representing the same semantic group with multiple images of the same object that would be suitable for the development and testing fine-grained recognition systems. In this respect, we believe that the PharmaPack objects acquired by modern mobile phones, under different acquisition conditions and on different backgrounds should fill this gap. Additionally, this dataset corresponds to a typical production chain of consumer goods that should be well suited for future scalability in mobile recognition applications.

B. State-of-the-art in mobile visual search and recognition

The generalized recognition system architecture under analysis is shown in Figure 1. The high dimensional feature extraction is based on either the usage of the last layers of deep nets trained in unsupervised or supervised way, a.k.a. *neural codes* [5], or aggregation of low- or- mid- dimensional local descriptors such as SIFT [6], SURF [7], aKaZe [8], etc. using feature aggregation such as Fisher vectors [9], VLAD [10], residual vectors [11], triangulated embedding [12], etc., that produces a resulting vector of defined length D. The resulting high-dimensional descriptors are collected in a database consisting of N enrolled feature vectors and the identification system should produce a list $\mathcal{L}(\mathbf{q})$ of indices of enrolled features $\mathbf{f}(i) \in \mathbb{R}^D$, $1 \leq i \leq N$ closest to the

ISBN 978-0-9928626-7-1 © EURASIP 2017

Fig. 2: Challenges in the fine-grained recognition of similar pharma products from PharmPack set.

Fig. 3: Challenges in recognition of authentic and fake packages according to [14].

probe feature vector $\mathbf{q} \in \mathbb{R}^D$. Additionally, local descriptors can be stored together with their coordinates within the image. In this case, one can explore the geometrical matching procedures between the enrolled descriptors and those of images to be verified. Such a matching is typically applied to a list of similar images returned based on an aggregated descriptor and it is referred to as *geometrical re-ranking* [13].

C. Particularities of pharma package recognition

Being a sub-task of the mobile visual search problem, the recognition of pharma packages is quite specific and has its own particularities that can be summarized as follows:

- *fine-grained recognition*: an accurate recognition of each unique pharma package is required in contrast to a similarity search based on approximate nearest neighbors (ANN) used in content retrieval systems. Many pharma packages are very similar to each other and the difference in appearance is really minor. While many methods based on advanced local descriptors and more recently on deep nets show very promising results on the recognition of distinctive classes (cars, people, animals, etc.) or in-class recognition (for example bird recognition which have very distinctive features), there are very little results on recognition of very similar objects such as those shown in Figure 2;
- *visual context*: it is not very rich and represents a mixture of text, logos and rarely some images. Text and graphical elements are very similar and local descriptors extracted from different packages are very close;
- *compactness of descriptors and memory footprint*: the descriptors should be very compact since there might be hundreds of millions of packages to recognize and the extracted features should be communicated via wireless networks to servers;
- *recognition conditions*: they are very varying due to light and geometry since the recognition is done using hand-held and mobile phones;
- *beyond recognition*: once the object is accurately recognized, we plan to decide whether it is authentic or not using special forensic features based on design accuracy of the fake package. This is shown in Figure 3 and will be referred to as *design verification*.

D. Contribution and objectives

In this paper, we try to cover the existing lack of modern datasets with a sufficient number of unique objects enrolled by mobile phones. We believe that the proposed dataset PharmaPack can be useful for many studies ranging from machine learning to security, especially in those applications requiring the fine-grained recognition and counterfeit detection. For future benchmarking, we present the first recognition results based on local descriptors with aggregation and compare them with the direct matching of local descriptors using geometrical information.

In particular, our objectives are: (1) To introduce the public database; (2) To give a fundamental estimation on the accuracy of recognition based on local descriptors next to geometric alignment based on RANSAC (here we do not consider any complexity issues); (3) To show how the local descriptors are suitable for fine-grained package recognition; (4) To show the impact of the number of local descriptors on the recognition accuracy; (5) To show the impact of acquisition conditions on recognition accuracy.

II. Problem formulation: PharmaPack dataset

A. Enrollment setup

We have collected 1000 unique packages. Each package was installed on a rotating table (covered in black color for better contrast) and 54 photos have been automatically taken by three mobile phones under different elevations and azimuths. We have used the fixed resolution of 8 Mpixels for all phones from Samsung. Moreover, following parameters were used: resolution - 3264x2448 (100 %); flash - off; AWB - auto; exposure compensation was -2.0 EV; ISO - auto (EV and ISO can be manually increased/decreased in rare cases for problematic/special packages); brightness - 0; sharpness - 0; saturation - 0; anti-banding mode(AB) - 50Hz. A typical example of 54 photos enrolled per one unique package is shown in Figure 4. Since the packages are of different sizes, we have assumed that the mobile recognition app will have a frame suggesting the end user to keep the package within this frame to avoid significant cropping and scaling. For this reason, we have enrolled packages from different distances depending on their size. The light conditions have been controlled by the external LEDs.

The local descriptors have been extracted from the package areas only, i.e., package areas have been automatically cropped from the acquired images. We extracted several types of local descriptors such as SIFT and aKaZe in two modes with a predefined number of descriptors (#desc. in figures) to be 300, 500 and 1000 and the varying number of descriptor set according to predefined threshold as in [15]. The public dataset PharmPack will provide both original and cropped images, all described extracted descriptors and their geometrical coordinates. The total number of descriptors extracted from all enrolled images is 175 Millions.

B. Recognition setup

In this work, we report the first results obtained for single object recognition in photo mode shooting on two backgrounds, namely, in fixed position on the surface and hand-held position on the same background as shown in

ISBN 978-0-9928626-7-1 © EURASIP 2017

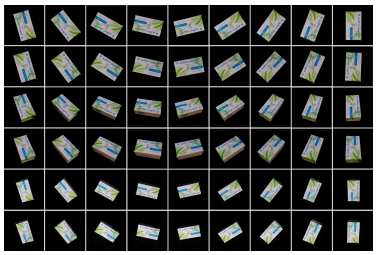


Fig. 4: Example of enrolled images.

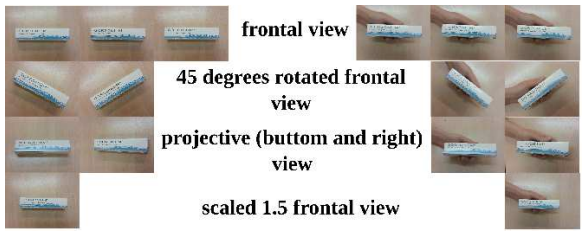


Fig. 5: Recognition datasets: examples of PharmaPack-R-I-S1 (left) and PharmaPack-R-I-S2 (right) acquisition.

Figure 5. These datasets are denoted as PharmaPack-R-I-S1 (S1) and PharmaPack-R-I-S2 (S2), respectively. Each recognition dataset contains 300 objects corresponding to the enrolled ones. A Samsung Galaxy S5 with 8 Mpixel resolution is used for the acquisition of 8 images per package corresponding to 3 frontal views, 2 45-degree-rotated frontal views, 2 projective views (right and bottom) and 1 1.5 scaled frontal view. In the recognition setup, the same phone parameters were used as in the enrollment except that the exposure compensation was set to 0.

III. Recognition methods under study

To investigate the impact of the number of descriptors, method of their aggregation and usefulness of geometrical information about the descriptors positions within the images, we have considered two setups that we will refer to as *aggregation setup* and *geometrical setup*. In all experiments for both datasets only grayscale images were used.

In the *aggregation setup*, the local descriptors have been aggregated in to Fisher Vectors (FV) following [9]. Gaussian Mixture Model (GMM) parameters were obtained from training data using the Expectation-Maximization algorithm [15]. In aggregation setup, SIFT and aKaZe descriptors extracted from 5400 enrolled images have been randomly chosen for training 128-, 256- and 512-component GMM.

In the *geometrical setup*, we have used RANSAC (RS) [16] for geometrical matching of distinctive descriptors [17].

Due to the computational burden, each image from the recognition sets was compared to 354 images from the Enrollment set, namely, to 54 images corresponded to the same object (see Figure 4) and to 300 images randomly chosen from dissimilar objects.

In the *aggregation setup*, we used the inner product to measure similarity of FVs. In the *geometrical setup*, we used the matching percentage from the total number of descriptors in the probe image.

Using these statistics, we have computed the ROC curves based on P_d and P_{fa} using the decision rule:

$$\begin{aligned} P_d &= \Pr\{S(i,j) \geq \gamma | \mathcal{H}_i\} \\ P_{fa} &= \Pr\{S(i,j) > \gamma | \mathcal{H}_{\bar{i}}\} \end{aligned}$$

where γ is the threshold and $S(i,j) = d(\mathbf{q}, \mathbf{f}_j(i))$ is a similarity measure between a probe feature vector $\mathbf{q}$ and an enrolled feature vector $\mathbf{f}_j(i)$, $\mathcal{H}_i$ and $\mathcal{H}_{\bar{i}}$ are correct and incorrect hypotheses respectively,

SIFT descriptors were extracted in two modes: (a) with a predefined number of descriptors to be 300, 500 and 1000 and (b) with varying number of descriptors chosen according to a defined reliability parameter, namely PeakThresh = 0.01. For aKaZe descriptors, we have not been able to determine the varying number of descriptors that would be suitable for all type of packages. Therefore aKaZe descriptors were extracted only for the same defined number of descriptors: 300, 500, 1000.

IV. Experimental results

Due to the limited space, we will restrict our results to the investigation of local descriptors based on popular aggregation methods such as Fisher Vectors and geometrical matching using geometrical coordinates of local descriptors based on RANSAC. We investigate several popular local descriptors such as SIFT considering it as a baseline and more recent one aKaZe due to the claimed superior performance for natural images and enhanced speed [8].

A. Recognition based on local descriptor aggregation

1) Impact of background: In order to investigate the impact of background, the descriptors were extracted from cropped and non-cropped images for all defined sets of parameters (for more details see Section III). For both datasets the obtained results for SIFT and aKaZe show the same effect, namely, the descriptors are localized in the regions of packages, but not on the background. Due to the lack of space, in Figure 6 only the results for feature matching of SIFT for the varying number of descriptors are shown. Since the background has small influence, all following results will be given only for cropped images.

2) Impact of recognition conditions and parameter selection: In order to investigate the impact of different GMM components, results for Fisher Vector matching of SIFT descriptors are obtained with respect to different numbers of GMM component (#comp. in figures), namely, 128, 256 and 512. Since the investigated dependencies are the same for both recognition datasets, in Figure 7 only results for the PharmaPack-R-I-S1 are reported. As illustrated in Figure 7, increasing the number of Gaussian components leads to an improvement of recognition accuracy. Although the 512-component Gaussian shows the best recognition accuracy, we decided to retain a 256-component Gaussian as it is less computationally expensive and very close in performance to the 512-component Gaussian.

In Figure 8, we present the obtained results for the matching of SIFT descriptors based on FV for both recognition datasets and both sets of parameters (defined and varying).

ISBN 978-0-9928626-7-1 © EURASIP 2017

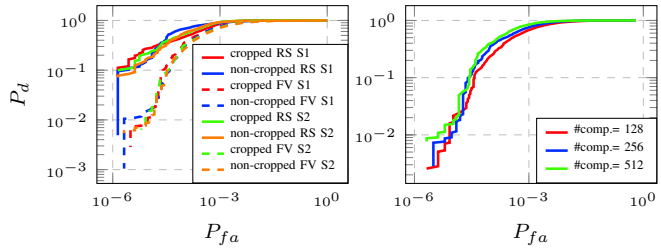


Fig. 6: Impact of background: cropped vs non-cropped based on RS and FV for SIFT descriptors(PeakThresh = 0.01, #comp. = 256).

Fig. 7: Impact of the number of GMM components in FV matching of SIFT for PharmaPack-R-I-S1 (PeakThresh=0.01).

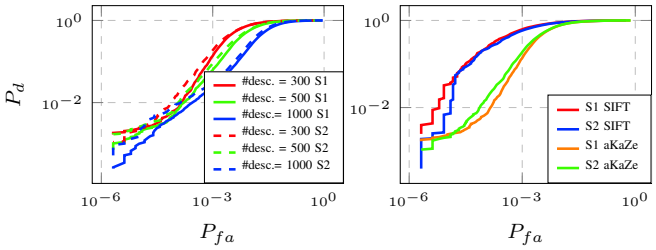


Fig. 9: Impact of recognition conditions and parameters of aKaZe descriptor based on FV.

Fig. 10: Impact of type of descriptors based on FV: SIFT vs aKaZe (#desc.= 300).

The recognition based on FV is used in list decoding mode. It should produce a very short list of candidates (list size $\sim P_{fa} \cdot N$, where N is the size of dataset) ensuring that the correct item is on the list with high probability. That is: $P_d \to 1$. According to the results reported in Figure 8, this leads to $P_{fa} \sim 10^{-2}$. Therefore, in a large scale database, i.e., when the number of enrolled items is around 10^9, the final short list will contain around 10^7 items which is far from being practical.

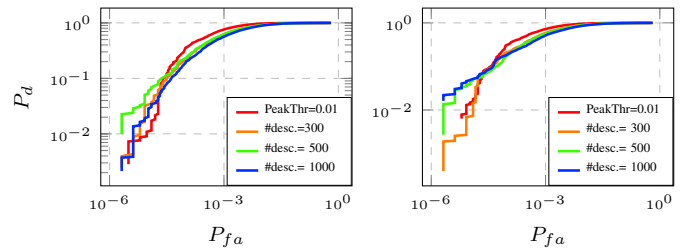
3) Impact of type of descriptors: In Figure 9, the results obtained for Fisher Vector matching of aKaZe descriptors for the predefined number of descriptors for both recognition datasets are illustrated. The results clearly demonstrate a pattern with regard to parameter selection in aKaZe features: increasing the number of descriptors leads to a decrease in recognition accuracy. The results for Fisher vector matching of SIFT and aKaZe descriptors illustrate different behavior with respect to the defined number of descriptors. In the case of SIFT descriptors (Figure 8), the best choice is the number of descriptors equal to 1000, whereas for aKaZe features the best result is obtained for a number of descriptors equal to 300. This is mainly due to the fact that when the number of aKaZe features increases, the feature points appear to be densely concentrated in local areas with a lot of overlap. In contrast to aKaZe, SIFT descriptors spread across the entire image. Therefore, from the point of view of aggregation, SIFT descriptors provide a more informative representation. In order to have a better comparison, the results of Fisher Vector matching of SIFT and aKaZe for 300 descriptors are shown in Figure 10. As expected, Figure 10 reveals that SIFT descriptors outperform aKaZe in the context of Fisher Vector matching for both datasets.

Summing up the results, it can be concluded that discarding the geometrical information leads to a loss of accuracy and inability of both SIFT and aKaZe to produce acceptable results for large scale systems.

B. Recognition based on RANSAC

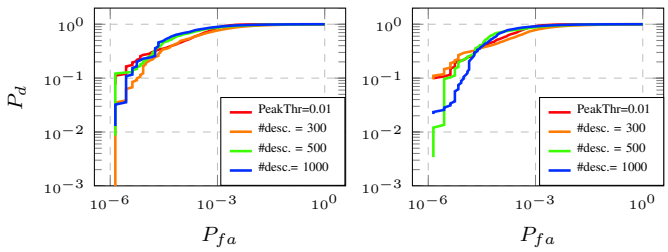
1) Impact of recognition conditions and parameter selection: In Figure 11, the results obtained for RANSAC matching of SIFT descriptors for the defined and varying number of descriptors for both recognition datasets are illustrated. First of all, it should be mentioned that the obtained recognition accuracy for both datasets is very similar but is a little bit higher for PharmaPack-R-I-S2. This is due to the fact that in PharmaPack-R-I-S1 the light condition is closer to those in the Enrollment set. Therefore, this is the reason behind the higher percentage of false positive matches amongst dissimilar packages like those represented in Figure 2. Because the packages enrolled for PharmaPack-R-I-S2 were hand-held during acquisition, the image quality is degraded. This subsequently causes a performance drop in matching. A subsequent side effect is that the number of false positives between different but visually near identical images, also drops, be it strictly due to the worse acquisition conditions. As for the parameter selection, in order to achieve $P_{fa} \sim 10^{-4} - 10^{-3}$, in both sets 1000 descriptors are needed. For smaller value of $P_{fa} \sim 10^{-6}$ in the case of PharmaPack-R-I-S1 the PeakThreshold = 0.01 is prefereable and the number of descriptors equal to 300 is better in the case of PharmaPack-R-I-S2.

2) Impact of type of descriptors: In Figure 12, the results obtained for RANSAC matching of aKaZe descriptors for the predefined number of descriptors for both datasets are illustrated. For PharmaPack-R-I-S2, it is a little bit better for the same reasons as in the case of SIFT. In contrast to the recognition based on FV, RANSAC for aKaZe performs



(a) PharmaPack-R-I-S1 (b) PharmaPack-R-I-S2

Fig. 8: Impact of recognition conditions and used parameters of SIFT based on FV.



(a) PharmaPack-R-I-S1 (b) PharmaPack-R-I-S2

Fig. 11: Impact of recognition conditions and parameters of SIFT based on RANSAC.

ISBN 978-0-9928626-7-1 © EURASIP 2017

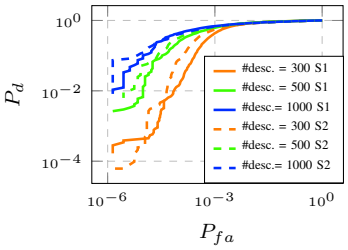


Fig. 12: Impact of recognition conditions and parameters of aKaZe based on RANSAC.

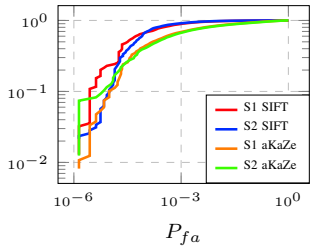


Fig. 13: Impact of type of descriptors based on RANSAC: SIFT vs aKaZe (#desc.= 1000).

better with increasing number of descriptors. However, it is true up to a certain limit after which the descriptors have excessive concentration in the same regions and do not add any useful information. Thus, we stopped at 1000 descriptors.

From Figures 11, 12, it is clear that SIFT is better than aKaZe for all used parameters except when #desc. equals 1000. For this parameter, the results of RANSAC matching for SIFT and aKaZe are shown in Figure 13. For PharmaPack-R-I-S1, SIFT is definitely the best. In the case of PharmaPack-R-I-S2, aKaZe demonstrates better results for P_{fa} less than 10^{-5}.

3) Recognition of similar objects: Unfortunately, the local descriptors are not suitable for fine-grained recognition. To demonstrate this, we use SIFT (#desc. = 1000, S1) for two similar but not identical packages as shown in Figure 14. It is easy to see that, from the point of view of local descriptors, these two packages can not be distinguished as dissimilar due to the big amount of matched descriptors. There are around 2K similar but not identical images in the database.

In conclusion, one can note that RANSAC based recognition with SIFT and aKaZe descriptors works well for distinctive objects. However, for fine-grained recognition both SIFT and aKaZe in RANSAC and FV recognition setups demonstrate unsatisfactory performance. It should be pointed out that the tested enrollment dataset is relatively small and includes only 1000 distinctive objects. In practice the targeted applications require perfect identification, that is $P_d = 1$. For this regime, both RANSAC and FV based recognition will retrieve around 50% of dataset. In a real scenario for a moderate dataset of 1000000 distinctive objects, the retrieved list will be no less than 500000 objects that is far too much for any practical system.

C. Public database

All raw labeled images, cropped images, extracted descriptors SIFT and aKaZe and aggregated descriptors using Fisher vectors will be available in the public domain upon paper acceptance at http://sip.unige.ch/pharmapack.

V. Conclusions

Undoubtedly, local descriptors are powerful tools for a wide range of tasks. In our experiments, we show the weak points of local descriptors such as SIFT and aKaZe from the point of view of fine-grained object recognition. For future work, we intend to investigate the recognition accuracy of a number of global descriptors such as GIST, descriptors produced by the last layers of deep networks trained on

Fig. 14: RANSAC matching based on SIFT (#desc. = 1000, S1): different packages. Around 40% of matched descriptors. The difference is in the number of pills and active components.

generic images and PharmaPack images and several descriptors specialized in text recognition. All datasets and results reported in this paper will be available in public domain to stimulate the reproducible research.

References

[1] J. M. Geusebroek, G. J. Burghouts, and A. W. M. Smeulders, "The Amsterdam library of object images," *International Journal of Computer Vision*, 2005.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[3] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results," http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[4] V. R. Chandrasekhar, D. M. Chen, S. S. Tsai, N.-M. Cheung, H. Chen, G. Takacs, Y. Reznik, R. Vedantham, R. Grzeszczuk, J. Bach, and B. Girod, "The stanford mobile visual search data set," in *Proceedings of the Second Annual ACM Conference on Multimedia Systems*. 2011, MMSys '11, pp. 117–122, ACM.

[5] A. Babenko, A Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. 2014, vol. 8689 of *Lecture Notes in Computer Science*, pp. 584–599, Springer.

[6] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the IEEE International Conference on Computer Vision*, 1999, vol. 2, pp. 1150–1157.

[7] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[8] Pablo F Alcantarilla and T Solutions, "Fast explicit diffusion for accelerated features in nonlinear scale spaces," *IEEE Trans. Patt. Anal. Mach. Intell*, vol. 34, no. 7, pp. 1281–1298, 2011.

[9] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, June 2007, pp. 1–8.

[10] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 3304–3311.

[11] Y. Chen, T. Guan, and C. Wang, "Approximate nearest neighbor search by residual vector quantization," *Sensors*, vol. 10, no. 12, pp. 11259–11273, 2010.

[12] H. Jégou and A. Zisserman, "Triangulation embedding and democratic aggregation for image search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3310–3317.

[13] H. Jegou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 117–128, Jan. 2011.

[14] "Fake vs real - fake black," `http://originalideas.info/easy-life/how-to-spot-a-fake-coco-chanel-mademoiselle`.

[15] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," `http://www.vlfeat.org/`, 2008.

[16] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[17] D. G Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

ISBN 978-0-9928626-7-1 © EURASIP 2017