

# LIGHTWEIGHT MOBILE OBJECT RECOGNITION WITH SEGMENTATION<sup>1</sup>

László Czúni, Péter József Kiss, Mónika Gál, Ágnes Lipovits, Metwally Rashad  
Image Processing Laboratory, University of Pannonia, Hungary Computer Science

**Abstract.** The purpose of our demo is to show the application and performance of some low-complexity image descriptors in object recognition under realistic circumstances. We built a client-server system where several image retrieval methods and image segmentation approaches can be tested with the help of a network connected Android device (mobile phone, tablet or head mounted computer). A modified version of the CEDD (Color and Edge Directivity Descriptor) is proposed, as the most robust lightweight descriptor found in our tests, and manual or automatic object segmentations are also included.

**Keywords:** object recognition, image distortion, color correction, image segmentation, mobile devices, client-server, demo

## 1 Introduction and motivation

With the availability of wearable and mobile computing devices automatic object recognition is an increasingly important research area. While some well-known methods gave significant breakthrough in the last 15 years [8] [3], due to the complexity of the real world and the great variability of environmental conditions there are still a lot of work to be done to create reliable, fast, low-complexity, and affordable recognition systems. In the recent years there is an increasing number of papers dealing with this topic, we just mention a few related to our work. A large set of approaches attempt to recognize simple (mostly binary) shapes by a mobile device (e.g. [5]). The purpose of these methods is to estimate camera pose and track objects to set an anchor for Augmented Reality applications. In [6] a simplified SIFT (Scale Invariant Feature Transform) and a scalable vocabulary tree is utilized for recognition. Mobile phone implementation aimed to recognize poster segments. In [4] in addition to the camera they used the accelerometer and the magnetic sensor to recognize the landscape. Clustered SURF (Speeded Up Robust Features) features were quantized using a vocabulary of visual words, learnt by k-means. For tracking objects the FAST corner detector was combined with sensor tracking. Because of the small storage capacity of the mobile device a server side service was needed to store the large number of images. [9] distinguishes the methods for the recognition of symbolic patterns (e.g. characters, shapes) and of natural objects (e.g. faces and flowers). It proposes, for the latter case, to apply user interactions. The image models are constructed automatically and corrected interactively only if necessary. It uses a mobile phone as a client and user interface to connect to the so called CAVIAR

---

<sup>1</sup> This paper is based on: László Czúni, Péter József Kiss, Ágnes Lipovits, Mónika Gál, Lightweight mobile object recognition, ICIP 2014, Paris

retrieval engine. In [12] a client-server application is introduced for multi-frame object detection. Tracking and feature extraction is implemented at the client side. Possible distortions are not investigated and only a few object classes are inspected. In [14] the problem of searching in large databases (several hundred thousand items) with mobile devices is attacked. The paper focuses on indexing (with bag of hash bits) and applies saliency based segmentation. It also states that drastic change in camera perspective and/or lighting, too small image/object size, non-rigid objects, insufficient (or non-discriminative) local features can cause serious problems in retrieval. In [13] no back-end server is used for processing. SIFT is used with hierarchical k-means classification to build a visual vocabulary on a mobile device. Unfortunately the robustness is not tested in a mobile imaging environment.

## **2 Implementation and usage**

We focus on a relatively general task where there are several views of a possibly 3D object, but the user approaches it basically from the front. Such can happen in several situations such as in interactive games; helping the daily life of the visually impaired; or in product recognition tasks such as getting information by pointing the camera at images in product catalogues. In a typical use-case people would like to use a wearable computer to recognize not more than a few hundred objects lying on the table or on a shelf. We do not use a 3D model of the objects but expect the system to recognize them from slight viewing angle deviations (about +/- 20 degree in each direction). Thus the reference database used for recognition is composed of these three views taken in normal lighting conditions. The demo has the ability to test several image descriptors via the help of a client-server application framework where the client is an Android based device and the server runs the different image analyzing algorithms. The most efficient descriptor (given later) is also implemented on the Android device and needs no remote server for operation. After the client takes an image, the region of interest can be selected manually (with cropping the target area of the object on the touch screen) or there is also an option to separate the object from its surrounding automatically based on the grabcut algorithm [15]. This later approach, favorable in case of head mounted computers, can separate foreground mostly in case homogeneously textured backgrounds. After successful recognition, the system displays textual meta (descriptive) information about the object. The client was implemented on Android 4.2 while the server runs on Windows Server 2012 with Windows Communication Foundations service. Running time from the request to displaying the best three matches is within 0.5 sec in case of about 100 trained objects in the database. The literature of object recognition is vast and we do not attempt to give a review in our limited length article. We just give a list of possible methods we thought would serve as the basis of a robust recognition engine. The methods tested for feature extraction and comparison can be grouped into four sets: MPEG-7 based methods [10] (MPEG7 CLD, MPEG7 EHD, MPEG7 SCD, MPEG7 Fusion); Local feature based methods (SURF, SURFVW [1], SIFT [8]); Compact Composite Descriptors [1] [2] (CompactCEDD, CEDD, CompactFCTH, FCTH, JCD, CCD Fusion, CompactVW); and others (Tamura texture descriptor, Color Correlogram and

Correlation (ACCC) [11], MPEG7-CCD Fusion [2]). For the detailed description of these algorithms please turn to the given references. Unfortunately, the SIFT based method ran extremely slow (about two orders slower) in initial tests compared to others and its performance was not better than the average of all so it was neglected in most of our tests and demo implementation. To select the most appropriate lightweight methods we built different test databases with hundreds of test images. The images, taken from 3 views, of different objects (such as shampoos, hardies and other cosmetic products, toys, office accessories) suffered different distortions such as different blur and motion blur effects, JPEG compression levels, additive noises, color distortions. These artifacts were added to reproduce the different image quality degradations that can happen in everyday life. During evaluation we measured the running time and average hit-rate (true positive rate) through running the query for all objects in the database (in this case there is no need to compute precision or accuracy due to equivalence). Contrary to the popularity of SIFT (and similar descriptors in its family such as SURF) in image retrieval we found serious drawbacks such as running time, sensitivity to blur and large storage requirements. Most promising methods were CEDD, FCTH, JCD, and CCD Fusion and the biggest loss in retrieval rate were caused by color balance distortions (hit-rate dropped to 8-64% from above 90%). For this reason we created a physical simulation of different lighting conditions in an experimental setup with the help of color tunable LED lamps. We reproduced 8 types of light sources: light tubes with 2700 and 4000 Kelvin; D50, D65, D70 standard light spectra; incandescent lamp, cold white and warm white. The spectral energy distribution of the light generated with LEDs was close to the aimed with negligible color differences. The reference images were taken under D65 with automatic white balance settings of the camera Canon EOS450D while the query images were made with two cameras with different settings: Canon G5 used automatic white balance (WB) but Canon EOS450 was fixed to D65. To improve the retrieval performance we selected the most compact size and fast CEDD and modified it by color normalization based on [7]. Color normalization increased the average performance from 22% to 58% in case of fixed white balance and from 46% to 82% in case of auto white balance in the lighting simulation experiments. CEDD [1] is a block based approach where each image block is classified into one of 6 texture classes with the help of MPEG7 EHD (Edge Histogram Descriptor). Then for each texture class a 24 bin color histogram is generated where each bin represents colors obtained by the division of the HSV color space. The values of the generated histogram of length 6x24 is then normalized and quantized to 8 bits. CEDD is one of the fastest methods with small descriptor size and showed quite robust behavior against distortions except for possible color balance problems.

### **3 Conclusion**

In this paper we showed a client-server application capable of fast and robust object recognition. As a major problem of successful object recognition color distortions were identified. After the comparison of several descriptors under different distortions

we found CEDD as one of the best lightweight methods. The proposed demo gives the ability to test the different retrieval methods and the manual or automatic segmentation of images of 3D objects.

## References

- [1] S. A. Chatzichristofis and Y. S. Boutalis: CEDD: Color and edge directivity descriptor A compact descriptor for image indexing and RETRIEVAL., 6th Int. Conf. in advanced research on Computer Vision Systems (ICVS), Lecture Notes in Computer Science (LNCS), pp.312–322, Santorini, Greece (2008)
- [2] S. A. Chatzichristofis, Y. S. Boutalis and M. Lux: Selection of the proper compact composite descriptor for improving content based image retrieval., The Sixth IASTED Int. Conf. on Signal Processing, Pattern Recognition and Applications (SPPRA), ACTA PRESS, pp.134–140, Innsbruck, Austria (2009)
- [3] G. Csurka, C. Dance, L.X. Fan, J. Willamowski, and C. Bray: Visual categorization with bags of keypoints?. Proc. of ECCV International Workshop on Statistical Learning in Computer Vision (2004)
- [4] S. Gammeter, A. Gassmann, L. Bossard, T. Quack, L. Van, Gool: Server-side object recognition and clientside object tracking for mobile augmented reality, IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1–8 (2010)
- [5] Hagbi, N., Bergig, O., El-Sana, J., Billinghamurst, M.: Shape Recognition and Pose Estimation for Mobile Augmented Reality, Mixed and Augmented Reality, 2009. ISMAR 2009. 8th IEEE International Symposium on, pp 65 – 71 ( 2009)
- [6] N. Henze, T. Schinke, and S. Boll: What is That? Object Recognition from Natural Features on a Mobile Phone, Proceedings of the Workshop on Mobile Interaction with the Real World, (2009)
- [7] N. Limare, J-L. Lisani, J-M. Morel, A. B. Petro, and C. Sbert: Simplest Color Balance, Image Processing On Line (2011)
- [8] D. G. Lowe: Distinctive Image Features from Scale-Invariant Keypoints. Int. J. Comput. Vision 60, 2 pp 91–110.(2004)
- [9] G. Nagy: Interactive, Mobile, Distributed Pattern Recognition, Proceedings of the International Conference on Image Analysis and Processing (ICIAP05), Lecture Notes in Computer Science, Springer, vol. LNCS 3617, pp. 37–49, Cagliari, (2005)
- [10] Sikora, T.:The MPEG-7 visual standard for content description-an overview,Circuits and Systems for Video Technology, IEEE Transactions on, vol.11, no.6, pp.696-702, (2001)
- [11] Tungkasthan, A., Intarasema, S., Premchaiswadi, W. Spatial Color Indexing using ACC Algorithm, ICT and Knowledge Engineering, pp.113–117, (2009)
- [12] Kumar, S.S. and Min Sun and Savarese, S.: Mobile object detection through client-server based vote transfer, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 3290–3297 (2012)
- [13] J. Panda, Michael S. Brown, C. V. Jawahar: Off-line Mobile Instance Retrieval with a Small Memory Footprint, IEEE International Conference on Computer Vision (ICCV), pp 1257–1264 (2013)
- [14] J. He, J. Feng, X. Liu, T. Cheng, T.-H. Lin, H. Chung, S.-F. Chang: Mobile Product Search with Bag of Hash Bits and Boundary Reranking, IEEE Conference on Computer Vision and Pattern Recognition CVPR (2012)
- [15] Rother, C., Kolmogorov, V., & Blake, A. (2004, August). Grabcut: Interactive foreground extraction using iterated graph cuts. In ACM Transactions on Graphics (TOG) (Vol. 23, No. 3, pp. 309-314). ACM.