Instant Scene Recognition on Mobile Platform

Sebastiano Battiato[1], Giovanni Maria Farinella[1],
Mirko Guarnera[2], Daniele Ravì[1], and Valeria Tomaselli[2]

[1] Image Processing Laboratory, University of Catania, Italy
{battiato,gfarinella,ravi}@dmi.unict.it
[2] Advanced System Technology, STMicroelectronics, Catania, Italy
{mirko.guarnera,valeria.tomaselli}@st.com

Abstract. Scene recognition is extremely useful to improve different tasks involved in the Image Generation Pipeline of single sensor mobile devices (e.g., white balancing, autoexposure, etc). This demo showcases our scene recognition engine implemented on a Nokia N900 smartphone. The engine exploits an image representation directly obtainable in the IGP of mobile devices. The demo works in realtime and it is able to discriminate among different classes of scenes. The framework is built by employing the FCam API to have an easy and precise control of the mobile digital camera. Each acquired image (or frame of a video) is holistically represented starting from the statistics collected on DCT domain. This allow instant and "free of charge" feature extraction process since the DCT is always computed into the IGP of a mobile for storage purposes (i.e., JPEG or MPEG format). A SVM classifier is used to perform the final inference about the context of the scene.

Keywords: Scene Recognition, DCT Features, FCam, Mobile Platform.

1 Introduction

Before shooting a photo, it is common practice to set focus, exposure and white balance taking into account the visual content of the observed scene. Clearly, a software engine able to automatically infer information about the category of a scene is extremely helpful to drive different tasks performed by single-sensor imaging devices during acquisition time (e.g., autofocus, autoExposure, white balance, etc.) or during post-acquisition time (e.g., image enhancement, image coding). For instance, the results reported in [1] demonstrate that the tuning of color constancy algorithms by taking into account the results of a scene classification engine (i.e., indoor vs outdoor) is useful to improve the quality of the final generated image. The need for the development of effective solution for scene recognition systems to be embedded in consumer imaging devices domain is confirmed by the growing interest of consumer devices industry which are including those capabilities in their products (e.g., Nikon, Canon, etc.). Different problems should be considered in transferring the ability of scene recognition into the IGP of a single imaging devices domain: memory limitation, low computational power, as well as the input data format to be used in scene recognition task (e.g., RAW, JPEG).

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part III, LNCS 7585, pp. 655–658, 2012.

In a recent work [2] we have proposed a scene recognition engine working in compressed and constrained domain. It exploits DCT features, fully compatible with the JPEG format (commonly used in consumer digital cameras). The decoding of the images is not necessary in extracting the features used to represent the scene. Local features are extracted by using simple operations directly in compressed DCT domain. Bank of Filters or extra information (eg., visual vocabulary) are not used during features extraction and image representation. The scenes are recognized at superordinate level of description (e.g., Natural vs Artificial, In vs Out, Open vs Closed) through a simple discriminative classifier. A very compact low dimensional vector of parameters is to be stored to perform final inference.

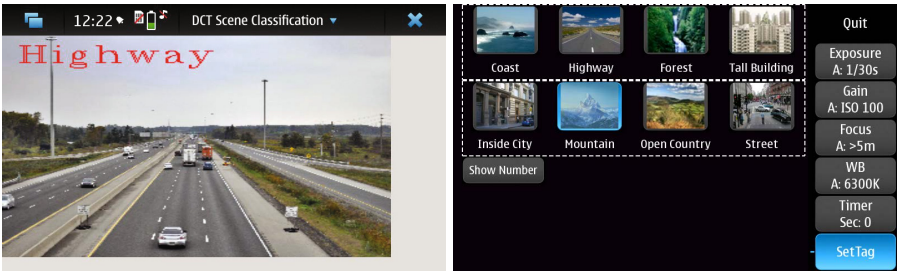
In [3] we have exploited distributions of textons on spatial hierarchy for scene classification at both, basic and superordinate level of description.

Building on [2,3] and considering the shape of the DCT coefficient distributions [4] to discriminate among different classes of scenes, this demo showcases our scene recognition engine implemented on a Nokia N900 smartphone. Although the limited resources of the hand-held device, the demo works in real-time and it is able to accurately discriminate among different classes of scenes. The features used for image representation can be retrieved directly in the IGP of a single sensor device. The outcome of the engine can be used to improve the different tasks involved into the IGP. The recognition results closely match state-of-the-art methods in terms of recognition accuracy.

2 Smartphone and Development Tools

Computational photography has been hampered, in the past, by the lack of a portable and programmable camera. In [5], a new open architecture and API for this kind of cameras has been introduced: the Frankencamera. One implementation of this architecture is embedded in the Nokia N900. This smartphone is equipped with a high-end OAP3430 ARM Cortex A8, running at 600 MHz, and runs the Maemo Linux distribution. The GPU is a PowerVR SGX 530, which supports OpenGL ES2.0. The system has 256 MB of dedicated high performance RAM (Mobile DDR) paired with access to 768 MB swap space managed by the OS, having a total of 1GB of virtual memory. The N900 has a five-megapixel Toshiba ET8EK8 image sensor, and it can capture full-resolution images at 12 frames per second, and VGA resolution at 30 frames per second. Since the Maemo OS is based on GNU Linux-kernel, developing for the N900 is similar to programming for any Linux device. Standard C++ code can be written using Qt, which creates applications that can be cross-compiled for different platforms. Programmers interact with the architecture by means of the "FCam" API, which allows to have access to the low level algorithms and data (e.g., lens position, exposure, gain, white balance, etc.). All the concepts of the FCam API (i.e., shots, sensors, frames and devices) are extensively described in [5,6].

The proposed demo implements our scene categorization method by processing the video stream grabbed by the device obtained through the "FCam" API.



(a) (b)

Fig. 1. (a) Recognition example of the engine implemented on the Nokia N900. (b) The GUI for tagging new images acquired with the Nokia N900.

The scene recognition engine is written exploiting also the OpenCV [7] and the LibSVM Libraries [8].

3 Demo and Beyond

We demonstrate a scene recognition engine operating on a Nokia N900 handheld device. The demo works in realtime and it is able to discriminate among 8 different classes of scenes at basic level of description: *Tall Building*, *Inside City*, *Street*, *Highway*, *Coast*, *Open Country*, *Mountain*, *Forest*.

We build on [2], where DCT features are used to build distributions of local dominant orientation (LDO) to discriminate at superordinate level of description. In this demo, rather than using LDO, the images are represented by encoding the shape of the DCT coefficients distributions [4] on a spatial hierarchy [3]. We consider only a subset of the 64 coefficients of each 8×8 DCT block of the image in order to capture both, edge features and textures. DCT information can be straight extracted during IGP of the device by allowing instant and "free of charge" feature extraction process since the 8×8 DCT representation of an image is always computed during the image generation for storage purposes (e.g., JPEG format). A SVM classifier is employed to perform the final inference about the context of the scene (Fig. 1(a)).

A user interface for tagging new images has been also realized to allow the user in populating the training dataset with new samples (e.g., for incremental learning purposes). Thanks to the FCam API, the tagging tool acquires not only images in their native CFA format (i.e., RAW), but also some useful statistics related to the shot (focus distance, estimated illumination temperature, exposure time, etc), that we are collecting to further improve the scene classification engine (Fig. 1(b)). The semantic categories of scenes that have been chosen for ground truth annotation are inspired to the ones proposed in [9].

4 Experimental Results

In table 1 we report the results obtained by our method on the 8 Scene Categories Dataset proposed by Oliva and Torralba in [9].

Table 1. Results obtained by our method on the 8 Scene Categories Dataset [9]

Confusion Matrix	*Tall Building*	*Inside City*	*Street*	*Highway*	*Coast*	*Open Country*	*Mountain*	*Forest*
Tall Building	0.88	0.07	0.00	0.01	0.01	0.00	0.01	0.02
Inside City	0.07	0.87	0.04	0.02	0.00	0.00	0.00	0.00
Street	0.03	0.04	0.89	0.02	0.00	0.01	0.01	0.01
Highway	0.00	0.03	0.02	0.82	0.07	0.03	0.03	0.00
Coast	0.00	0.00	0.00	0.02	0.85	0.11	0.01	0.01
Open Country	0.00	0.00	0.01	0.02	0.15	0.74	0.05	0.03
Mountain	0.01	0.00	0.00	0.01	0.02	0.05	0.85	0.06
Forest	0.00	0.00	0.00	0.00	0.00	0.02	0.05	0.93

The recognition results closely match state-of-the-art methods in terms of recognition accuracy. The results in Table 1 can be compared with respect to the ones obtained by employing the GIST descriptor [9] reported at the following URL: http://people.csail.mit.edu/torralba/code/spatialenvelope/. One should not overlook that our recognition engine has no computational overhead for feature extraction since it works with simple features that are directly computed in the IGP of mobile single sensor devices. This highly reduce the complexity of the scene recognition system.

Acknowledgments. The authors would like to thanks Nokia Research Center Palo Alto for providing the N900 smartphones.

References

1. Bianco, S., Ciocca, G., Cusano, C., Schettini, R.: Improving color constancy using indoor - outdoor image classification. IEEE Transactions on Image Processing 17, 2381–2392 (2008)
2. Farinella, G.M., Battiato, S.: Scene classification in compressed and constrained domain. IET Computer Vision 5, 320–334 (2011)
3. Battiato, S., Farinella, G.M., Gallo, G., Ravì, D.: Exploiting textons distributions on spatial hierarchy for scene classification. Journal on Image and Video Processing 2010, 7:1–7:13 (2010)
4. Lam, E., Goodman, J.: A mathematical analysis of the dct coefficient distributions for images. IEEE Transactions on Image Processing 9, 1661–1666 (2000)
5. Adams, A., Talvala, E.V., Park, S.H., Jacobs, D.E., Ajdin, B., Gelfand, N., Dolson, J., Vaquero, D., Baek, J., Tico, M., Lensch, H.P.A., Matusik, W., Pulli, K., Horowitz, M., Levoy, M.: The frankencamera: an experimental platform for computational photography. ACM Transactions on Graphics - Proceedings of ACM SIGGRAPH 2010 29, 29:1–29:12 (2010)
6. FCam Garage: FCam API (2012), `http://fcam.garage.maemo.org/`
7. Willow Garage: OpenCV: Open source computer vision library (2012), `http://opencv.willowgarage.com/wiki/`
8. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines (2012), `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`
9. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. International Journal of Computer Vision 42, 145–175 (2001)