Pattern Recognition 48 (2015) 1086–1100

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Representing scenes for real-time context classification on mobile devices

G.M. Farinella [a,*], D. Ravì [a], V. Tomaselli [b], M. Guarnera [b], S. Battiato [a]

[a] Image Processing Laboratory, University of Catania, Italy
[b] Advanced System Technology – Computer Vision, STMicroelectronics, Catania, Italy

ARTICLE INFO

Article history:
Received 30 November 2013
Received in revised form
13 May 2014
Accepted 19 May 2014
Available online 9 June 2014

Keywords:
Scene representation
Scene classification
Image descriptor
GIST
JPEG
DCT features
Mobile devices
Wearable cameras

ABSTRACT

In this paper we introduce the DCT-GIST image representation model which is useful to summarize the context of the scene. The proposed image descriptor addresses the problem of real-time scene context classification on devices with limited memory and low computational resources (e.g., mobile and other single sensor devices such as wearable cameras). Images are holistically represented starting from the statistics collected in the Discrete Cosine Transform (DCT) domain. Since the DCT coefficients are usually computed within the digital signal processor for the JPEG conversion/storage, the proposed solution allows to obtain an instant and "free of charge" image signature. The novel image representation exploits the DCT coefficients of natural images by modelling them as Laplacian distributions which are summarized by the scale parameter in order to capture the context of the scene. Only discriminative DCT frequencies corresponding to edges and textures are retained to build the descriptor of the image. A spatial hierarchy approach allows to collect the DCT statistics on image sub-regions to better encode the spatial envelope of the scene. The proposed image descriptor is coupled with a Support Vector Machine classifier for context recognition purpose. Experiments on the well-known 8 Scene Context Dataset as well as on the MIT-67 Indoor Scene dataset demonstrate that the proposed representation technique achieves better results with respect to the popular GIST descriptor, outperforming this last representation also in terms of computational costs. Moreover, the experiments pointed out that the proposed representation model closely matches other state-of-the-art methods based on bag of Textons collected on spatial hierarchy.

1. Introduction and motivations

Scene recognition is a key process of human vision which is exploited to efficiently and rapidly understand the context and objects in front of us. Humans are able to recognize complex visual scenes at a single glance, despite the number of objects with different poses, colors, shadows and textures that may be contained in the scenes. Seminal studies in computational vision [1] have portrayed scene recognition as a progressive reconstruction of the input from local measurements (e.g., edges and surfaces). In contrast, some experimental studies have suggested that recognition of real-world scenes may be initiated from the encoding of the global configuration, bypassing most of the details about local concepts and objects information [2]. This ability is achieved mainly by exploiting the holistic cues of scenes that can be processed as single entity over the entire human visual field without requiring attention to local features [3]. Successive studies suggest that the humans rely on local as much as on global information to recognize the scene category [4,5].

The recognition of the scene is a useful task for many relevant Computer Vision applications: robot navigation systems [6], semantic organization of databases of digital pictures [7], content-based image retrieval (CBIR) [8], context driven focus attention and object priming [9,10], and scene depths estimation [11]. To build a scene recognition system, different considerations about the spatial envelope properties (e.g., degree of naturalness and degree of openness) and the level of description of the scene (e.g., subordinate, basic, and superordinate) have to be taken into account [12].

The results reported in [13] demonstrate that a context recognition engine is important for the tuning of color constancy algorithms used in the Imaging Generation Pipeline (IGP) and hence improve the quality of the final generated image. More in general, in the research area of single sensor imaging devices [14], the scene context information can be used to drive different tasks performed in the IGP during both acquisition time (e.g., autofocus, auto-exposure, and white balance) and post-acquisition time (e.g.,

* Corresponding author. Tel.: +39 3477965844; fax: +39 095330094.
E-mail addresses: gfarinella@dmi.unict.it (G.M. Farinella), ravi@dmi.unict.it (D. Ravì), valeria.tomaselli@st.com (V. Tomaselli), mirko.guarnera@st.com (M. Guarnera), battiato@dmi.unict.it (S. Battiato).

http://dx.doi.org/10.1016/j.patcog.2014.05.014
0031-3203/

image enhancement and image coding). For example, the auto-scene mode of consumer and wearable cameras could allow to automatically set the acquisition parameters improving the perceived quality of the captured image according to the recognized scene (e.g., Landscape and Portrait). Furthermore, context recognition could be functional for the automatic setting of surveillance cameras which are usually placed in different scene contexts (e.g., Indoor vs Outdoor scenes and Open vs Closed scenes), as well as in the application domain of assistive technologies for visually impaired and blind people (e.g., indoor vs outdoor recognition with wearable smart glasses). The need for the development of effective solution for scene recognition systems to be embedded in consumer imaging devices (e.g., consumer digital cameras, smartphones, and wearable cameras) is confirmed by the growing interest of consumer devices industry which are including those capabilities in their products. Different constraints have to be considered in transferring the ability of scene recognition into the IGP of a single sensor imaging devices [15]: memory limitation, low computational power, as well as the input data format to be used in scene recognition task (e.g., JPEG images).

This paper presents a new computational model to represent the context of the scene based on the image statistics collected in the Discrete Cosine Transform (DCT) domain. We call DCT-GIST the proposed scene context descriptor. Since the DCT of the image acquired by a device is always computed for JPEG conversion/storage,[1] the features extraction process useful to compute the signature of the scene context is "free of charge" for the IGP and can be performed in real-time independently from the computational power of the device. The rationale beyond the proposed image representation is that the distributions of the AC DCT coefficients (with respect to the different AC DCT basis) differ from one class of scene context to another and hence can be used to discriminate the context of scenes. The statistics of the AC DCT coefficients can be approximated by a Laplacian distribution [16] almost centered at zero; we extract an image signature which encodes the statistics of the scene by considering the scales of Laplacian models fitted over the distribution of AC DCT coefficients of the image under consideration (see Fig. 1). This signature computed on a spatial pyramid [17,18], together with the information related to the colors obtained considering the DC components, is then used for the automatic scene context categorization.

To reduce the computational complexity involved in the image representation extraction, only a subset of the DCT frequencies (summarizing edges and textures) are considered. To this purpose a supervised greedy based selection of the most discriminative frequencies is performed. To improve the discrimination power, the spatial envelope of the scene is encoded with a spatial hierarchy approach useful to collect the AC DCT statistics on image sub-regions [17,18]. We have coupled the proposed image representation with a Support Vector Machine classifier for final context recognition purpose. The experiments performed on the 8 Scene Context Dataset [12] as well as on the MIT-67 Indoor Scene dataset [5] demonstrate that the proposed DCT-GIST representation achieves better results with respect to the popular GIST scene descriptor [12]. Moreover, the novel image signature outperforms GIST in terms of computational costs. Finally, with the proposed image descriptor we obtain results comparable with other more complex state-of-the-art methods exploiting spatial pyramids [17] and combination of global and local information [5].

The primary contribution of this work is related to the new descriptor for scene context classification which we call DCT-GIST. We emphasize once again the fact that the proposed descriptor is built on information already available in the IGP of single sensor devices as well as in any image coded in JPEG format. Compared to many other scene descriptors extracted starting from RGB images [4,12,13,17–20], the proposed representation model has the following peculiarities/advantages:

- the decoding/decompression of JPEG is no needed to extract the scene signature;
- visual vocabularies have not to be computed and maintained in memory to represent both training and test images;
- the extraction of the scene descriptor does not need complex operation such as convolutions with bank of filters or domain transformations (e.g., FFT);
- there is no need of a supervised/unsupervised learning process to build the scene descriptor (e.g., there is no need of pre-labeled data and/or clustering procedure);
- it can be extracted directly into the Imaging Generation Pipeline of mobile devices with low computational resources;
- the recognition results closely match state-of-the-art methods cutting down the computational resources (e.g., computational time needed to compute the image representation).

The remainder of this paper is organized as follows: Section 2 briefly surveys the related works. Section 3 gives the background about the AC DCT coefficients distributions for different image categories. Section 4 presents the proposed image representation, whereas the new Image Generation Pipeline architecture is described in Section 5. Section 6 reports the details about the experimental settings and discusses the obtained results. Finally, Section 7 concludes the paper with hints for future works.

2. Related works

The visual content of the scene can be described with local or global representation models. A local based representation of the image usually describes the context of the scene as a collection of previously recognized objects/concepts within the scene, whereas a global (or holistic) representation of the scene context considers the scene as a single entity, bypassing the recognition of the constituting concepts (e.g., objects) in the final representation. The representation models can significantly differ for their capability of extracting and representing important information for the context description.

Many Computer Vision researchers have proved that holistic approaches can be effectively used to solve the problem of rapid and automatic context recognition. Most of the holistic approaches share the same basic structure that can be schematically summarized as follows:

1. A suitable features space is considered (e.g., textons vocabularies [17]). This space must emphasize specific image cues such as corners, oriented edges, and textures.
2. Each image under consideration is projected into the considered feature space. A descriptor is built considering the image as a whole entity (e.g., textons distributions [17]).
3. Context recognition is obtained by using Pattern Recognition and Machine Learning algorithms on the computed representation of the images (e.g., by using K-nearest neighbors and SVM).

A wide class of techniques based on the above scheme, works extracting features on perceptually uniform color spaces (e.g., CIELab). Typically, filter banks [19,21] or local invariant descriptors [18,20] are employed to capture image cues and to build the visual vocabulary to be used in a bag of visual words model [22]. An image is considered as a distribution of visual words and this

[1] JPEG is the most common used format for images and videos.