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.

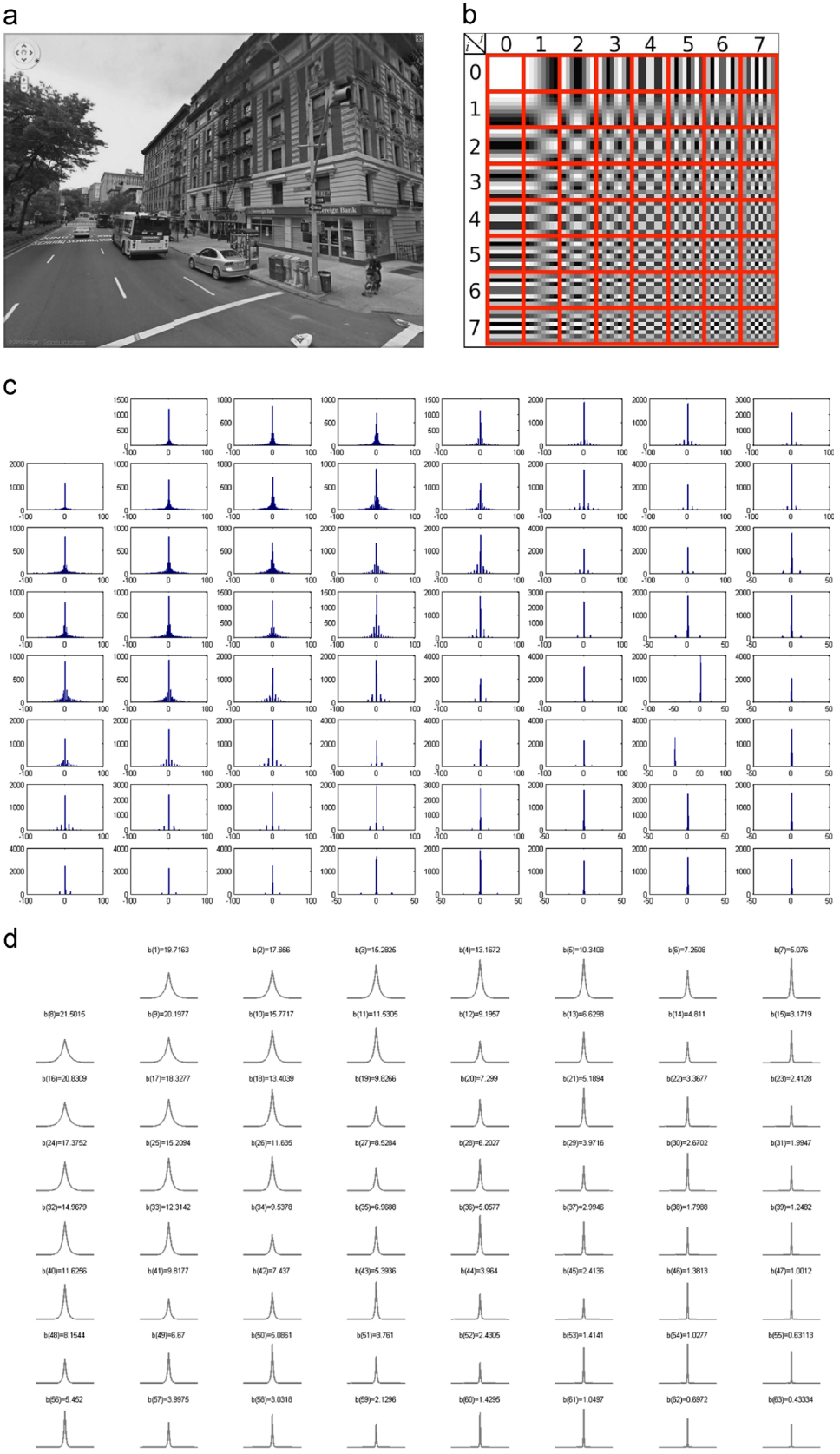

Fig. 1. Given the luminance channel of an image (a), the feature vector associated to the context of the scene is obtained considering the statistics of the AC coefficients corresponding to the different AC DCT basis (b). For each AC frequency, the coefficients distribution is computed (c) and fitted with a Laplacian model (d). Each fitted Laplacian is characterized by a scale parameter related to the slope of the distribution. The final image signature is obtained collecting the scale parameters of the fitted Laplacians among the different AC DCT coefficient distributions. As specified in Section 4, information on colors (i.e., DC components) as well as on the spatial arrangement of the DCT feature can be included to obtain a more discriminative representation. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

holistic representation is used for classification purposes. Spatial information have been also exploited in order to capture the layout of the visual words within images [18,23]. A review of some other state-of-the-art methods working with features extracted on spatial domain can be found in [24].

On the other hand, different approaches have considered the frequency domain as an useful and effective source of information to holistically encode an image for scene classification. The statistics of natural images on frequency domain reveal that there are different spectral signatures for different image categories [25]. In particular by considering the shape of the FFT spectrum of an image it is possible to address scene category [12,25,26], scene depth [11], and object priming such as identity, scale and location [10].

As suggested by different studies in computational vision, scene recognition may be initiated from the encoding of the global configuration of the scene, disregarding details and object information. Inspired by this knowledge, Torralba and Oliva [26] have introduced computational procedures to extract the global structural information of complex natural scenes looking at the frequency domain [12,25,26]. The computational model presented in [26] works in the Fourier domain where Discriminant Structural Templates (DSTs) are built using the power spectrum. A DST is a weighting scheme over the power spectrum that assigns positive values to the frequencies that are representative for one class and negative for the others. In particular the sign of the DST values indicates the correlation between the spectral components and the "spatial envelope" properties of the two groups to be distinguished. When the task is to discriminate between two kinds of scenes (e.g., *Natural* vs. *Artificial*) a suitable DST is built and used for the classification. A DST is learned in a supervised way using Linear Discriminant Analysis. The classification of a new image is hence performed by the sign of the correlation between the power spectrum of the considered image and the DST. A relevant issue in building a DST is the sampling of the power spectrum both at the learning and classification stages (a bank of Gabor filters with different frequencies and orientation is used in [26]). The final classification is performed on the Principal Components of the sampled frequencies. The improved version of the DST descriptor is called GIST [12,25]. Oliva and Torralba [12] performed tests using GIST on a dataset containing pictures of 8 different environmental scenes covering a large variety of outdoor places obtaining good performances. The GIST descriptor is nowadays one of the most used representation to encode the scene as whole. It has been used in many Computer Vision application domains such as robot navigation [6], visual interestingness [27], image retrieval [28], and video summarization [29].

Luo and Boutell [30] built on previous works of Torralba and Oliva [26] and proposed to use Independent Component Analysis rather than PCA for features extraction. In addition they have combined the camera metadata related to the image capture conditions with the information provided by the power spectra to perform the final classification.

Farinella et al. [31] proposed to exploit features extracted by ordering the Discrete Fourier Power Spectra (DFPS) to capture the *naturalness* of scenes. By ordering the DFPS the overall "shape" of the scene in frequency domain is captured. In particular the frequencies that better capture the differences in the energy "shapes" related to *Natural* and *Artificial* categories are selected and ordered by their response values in the Discrete Fourier power spectrum. In this way a "ranking number" (corresponding to the relative position in the ordering) is assigned to each discriminative frequency. The vector of the response values and the vector of the relative positions in the ordering of the discriminative frequencies are then used singularly or in combination to provide a holistic representation of the scene. The representation was used with a probabilistic model for *Natural* vs *Artificial* scene classification.

The Discrete Cosine Transform (DCT) domain was explored by Farinella and Battiato [15] to build histograms of local dominant orientations to be used as scene representation at the abstract level of description (e.g., *Natural* vs *Artificial* and *Indoor* vs *Outdoor*). The representation is built collecting the information about orientation and strength of the edges related to the JPEG image blocks [7]. This representation was coupled with a logistic classifier to discriminate between the different scene contexts.

The aforementioned techniques disregard the spatial layout of the discriminative frequencies. Seminal studies proposed by Torralba et al. [9–11] have proposed to further look at the spatial frequency layout to address more specific vision tasks by exploiting contextual information (e.g., object detection and recognition, and scene depth estimation).

3. The statistics of natural image categories in DCT domain

One of the most popular standard for lossy compression of images is the JPEG [32]. The JPEG compression is available in every IGP of single sensor consumer devices such as digital consumer cameras, smartphones and wearable cameras (e.g., smart glasses). Moreover, most of the images on Internet (e.g., in social networks and websites) are stored in JPEG format. Nowadays, around 70% of the total images on the top 10 million websites are in JPEG format.[2] Taking into account these facts, a scene context descriptor that can be efficiently extracted in the IGP and/or directly in the JPEG compressed domain is desirable.

The JPEG algorithm splits the image into non-overlapping blocks of size 8×8 pixels and each block is then processed with the Discrete Cosine Transform (DCT) before quantization and entropy coding [32]. The DCT has been studied by many researchers which have proposed different models for the distributions of the DCT coefficients. One of the first conjecture was that the AC coefficients have Gaussian distributions [33]. Different other possible distributions of the coefficients have also been proposed, including Cauchy, generalized Gaussian, as well as a sum of Gaussians [34–38]. The knowledge about the mathematical form of the statistical distribution of the DCT coefficient is useful for the quantizer design and noise mitigation for image enhancement. Although methods to extract features directly from JPEG compressed domain have been presented in the literature in the application context of image retrieval [39,40], at the best of our knowledge there are not works where the DCT coefficients distributions are exploited for scene classification. The proposed image representation is inspired by the works of Lam [16,41], where the semantic content of the images has been characterized in terms of DCT distributions modelled as Laplacian and generalized Gaussian models.

After performing the DCT on each 8×8 block of an image and collecting the corresponding coefficients to the different AC basis of the DCT, a simple observation of the distribution indicates that they resemble a Laplacian (see Fig. 1(c)). This guess has been demonstrated through a rigorous mathematical analysis in [16]. The probability density function of a Laplacian distribution can be written as

$$f(x|\mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right) \quad (1)$$

where μ is the location parameter and $b \geq 0$ is the scale parameter. Fig. 2 reports examples of different Laplacian distributions. At varying of the scale parameter, the Laplacian distribution changes

[2] Source: http://w3techs.com/technologies/overview/image_format/all. The statistics is computed on the top 10 million websites according to the Amazon.com company (Nov 2013).

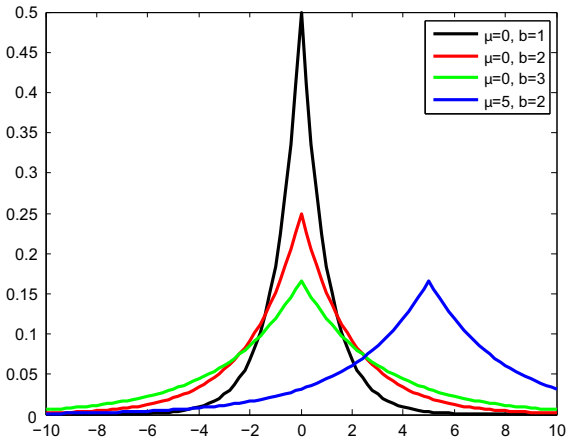

Fig. 2. Laplacian distribution at varying of μ and b.

its shape. Given N samples $\{x_1,\ldots,x_N\}$, the parameters μ and b can be simply estimated with the maximum likelihood estimator [42]. Specifically, μ corresponds to the median of the samples,[3] whereas b is computed as follows:

$$b = \frac{1}{N}\sum_{i=1}^{N}|x_i-\mu|. \quad (2)$$

The rationale beyond the proposed representation for scene context classification is that the context of different classes of scenes differs in the scales of the AC DCT coefficient distributions. Hence, to represent the context of the scene we can use the feature vector of the scales of the AC DCT coefficients distributions of an image after a Laplacian fitting. Fig. 3 reports the average "shapes" of the AC DCT coefficients Laplacian distributions related to the 8 Scene Context Dataset [12]. The dataset contains 2600 color images (256×256 pixels) belonging to the following 8 outdoor scene categories: *coast*, *mountain*, *forest*, *open country*, *street*, *inside city*, *tall buildings*, *highways*. The Laplacian shapes in Fig. 3 are computed by fitting the Laplacian distributions for the different AC DCT coefficients of the luminance channel of each image and then averaging the Laplacian parameters with respect to the 8 different classes (color coded in Fig. 3). A simple observation of the slopes of the different Laplacian distributions (corresponding to the b parameter) is useful to better understand the rationale beyond the proposed scene descriptor. The slopes related to the different classes are captured by the b parameters computed (with low computational cost) from the images directly encoded in the DCT domain (i.e., JPEG format). The guess is that the multidimensional space of the b parameters is discriminative enough for scene context recognition. Although it is difficult to visualize the N-dimensional distributions of the b parameters, an intuition of the discriminativeness of the space can be obtained by considering two AC DCT frequencies and plotting the 2-dimensional distributions of the related Laplacian parameters. Fig. 4 shows the 2-dimensional distributions obtained by considering two DCT frequencies corresponding to the DCT basis (0,1) and (1,0) which are useful to reconstruct the vertical/horizontal edges of each image block (see Fig. 1(b)). As the figure points out, already considering only two AC DCT frequencies there is a good separation among the eight different classes. The experiments reported in Section 6 quantitatively confirm the above rationale.

[3] Note that for the different AC DCT distributions the μ value is not equal to zero.

4. Proposed DCT-GIST image representation

In this section we formalize the proposed image representation which builds on the main rationale that different scene classes have different AC DCT coefficient distributions (see Section 3). Fig. 3 shows the average of the AC DCT coefficient distributions after a Laplacian fitting on images belonging to different scene contexts. Differences in the slopes of the Laplacian distributions are evident and related to the different classes. As a consequence of this observation, we propose to encode the scene context by concatenating all the Laplacian parameters related to the median and slope (μ and b) which are computed by considering the different AC DCT coefficients distributions of the luminance channel of the image.[4] In addition to these information, the mean and variance of the DC coefficients can be also included into the feature vector to capture the color information, as well as the AC DCT Laplacian distributions parameters obtained considering the C_b and C_r channels.[5] In Section 6 we show the contribution of each component involved in the proposed DCT-GIST image descriptor.

The aforementioned image features are extracted in the IGP just after the image acquisition step, without any extra complex processing. Specifically, the Laplacian parameters related to the AC DCT coefficients are obtained collecting the AC DCT coefficients inside the JPEG encoding module performed before the image storage. In case the image is already stored in JPEG (e.g., a picture from the web), the information useful for scene context representation can be directly collected in the compressed domain without any further processing. Indeed, to build the scene descriptor in the DCT compressed domain, only simple operations (i.e., the median and the mean absolute deviations from the median) are needed to compute μ and b for the different image channels, as well as to compute the mean and variance on the DC components. This cuts down the computational complexity with respect to other descriptors which usually involve convolution operations (e.g., with bank of filters [17] or Gaussian Kernels [12]) or other more complex pipelines (e.g., Bag of Words representation [18]) to build the final scene context representation.

It is well-known that some of the DCT basis are related to the reconstruction of the edges of an 8×8 image block (i.e., first row and first column of Fig. 1(b)), whereas the others are more related to the reconstruction of the textured blocks. As shown in [43] the most prominent patterns composing natural images are the edges. High frequencies are usually affected by noise and could be not really useful for discriminating the context of a scene. For this reason we have performed an analysis to understand which of the AC DCT basis can give a real contribution to discriminate between different classes of scenes. One more motivation to select only the most discriminative AC DCT frequencies is the reduction of the complexity of the overall system.

To properly select the AC DCT frequencies to be employed in the final image representation, we have collected (from Flickr) and labelled a set of 847 uncompressed images to be used as validation set. These images belong to the eight different classes of scene context [12] (see Fig. 3) and have variable size (max size 6000×4000, min size 800×600). We used uncompressed images to avoid that the selection processes of the most discriminative frequencies could be biased by the JPEG quantization step. On this dataset we have performed scene context classification by representing images through the Laplacian fitting of a single AC DCT basis. This step has been repeated for each AC DCT basis. A greedy fashion approach has been hence employed to select the most discriminative frequencies. This means that as first round the classification has been performed for all the AC DCT basis

[4] Note that in the JPEG format the image is converted in the YC_bC_r color model as first step.

[5] The DCT chrominance exhibits the same distribution as for the luminance channel [16].

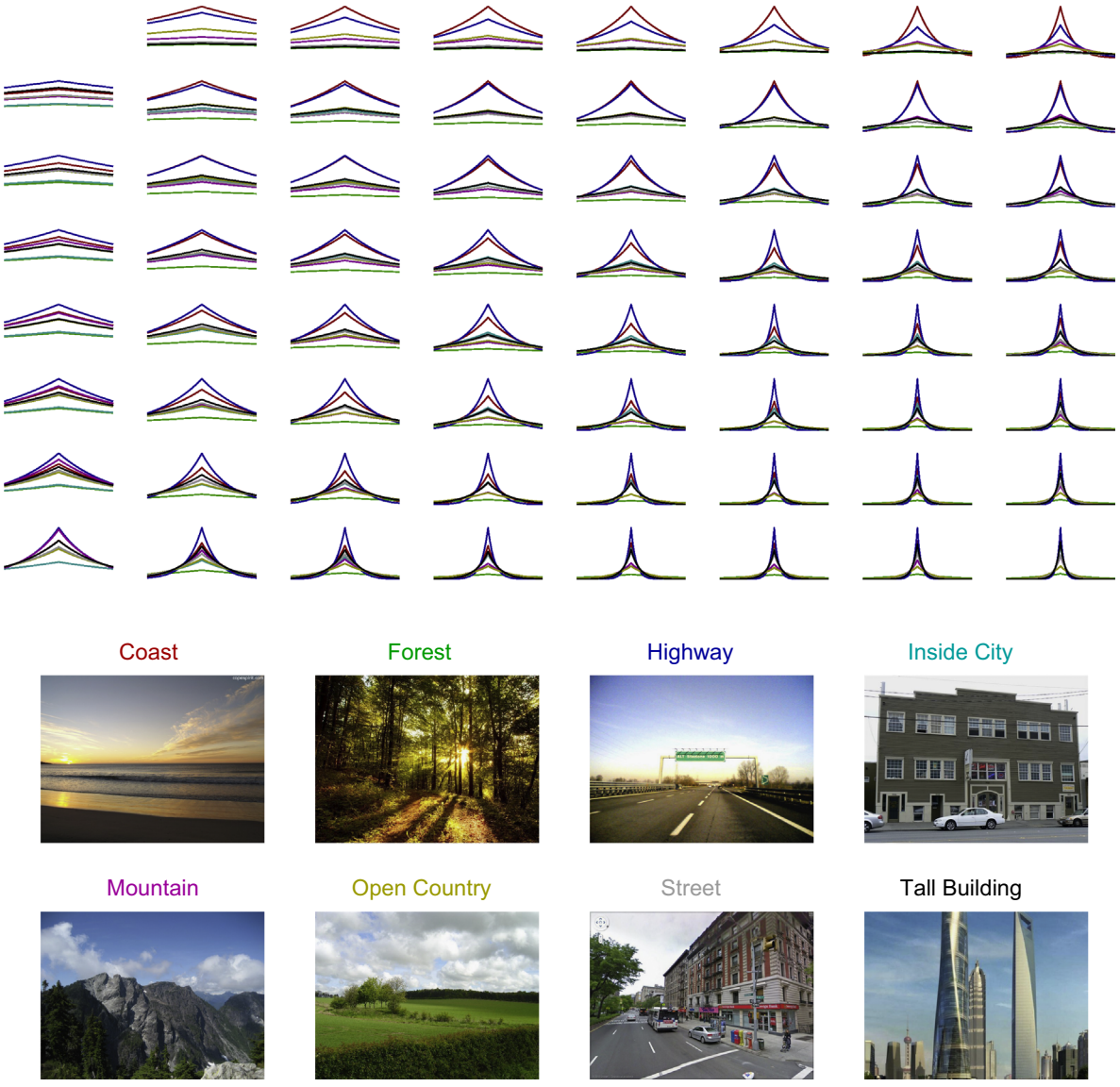

Fig. 3. Average Laplacian distributions of the AC DCT coefficients considering the 8 Scene Context Dataset [12]. The different scene classes are color coded. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

separately. The images have been hence classified after performing the learning of a support vector machine. A leave one out modality has been used to evaluate the discriminativeness of each AC DCT basis. Then we have selected the most discriminative frequency and we have performed another round of learning and classification considering the selected frequency coupled with one of the remaining frequencies in order to jointly consider two AC DCT basis. This procedure has been recursively repeated to greedily select frequencies. The experiments on the validation set suggested that a good trade-off between context classification accuracy and computational complexity (i.e., the number of AC DCT frequencies to be included in a real IGP to fit with required computational time and memory resources) is the one which considers the AC DCT frequencies marked in red in Fig. 5. Let $D(i,j)$, $i=1,\ldots,7$, $j=1,\ldots,7$, be the DCT components corresponding to the 2D DCT basis (i,j) in Fig. 1(b). The final set of the selected AC DCT basis in Fig. 5 is defined as

$$F=\{(i,0)|i=1,\ldots,7\}\bigcup\{(i,1)|i=1,\ldots,3\}\bigcup\{(0,j)|j=1,\ldots,7\}\bigcup\{(1,j)|j=1,\ldots,3\}\bigcup\{(i,j)|i=0\ldots,7;j=7-i\}. \quad (3)$$

Table 1 reports the accuracy obtained on the aforementioned validation dataset considering the Laplacian fitting of all the 63 AC DCT basis, as well as the results obtained considering the 25 selected basis in Eq. (3) (see Fig. 5). Notice that the overall accuracy obtained with the only 25 selected AC DCT basis is higher than the one obtained by considering all the 63 AC DCT basis. This is due to the fact that high frequencies (i.e., the ones below the diagonal in Fig. 5) could contain more noise information than the other frequencies, making confusion into the feature space.

The scene context descriptor proposed so far, uses a global feature vector for describing an image by leaving out the information about the spatial layout of the local features. The relative position of a local descriptor can help to disambiguate concepts that are similar in terms of local descriptor. For instance, the visual concepts “sky” and “sea” could be similar in terms of local descriptor, but they are typically different in terms of position within the scene. The relative position can be thought as the context in which a feature takes part with respect to the other features within an image. To encode information of the spatial layout of the scene, different pooling strategies have been proposed in literature [17,18]. Building on our previous work [17] we have augmented the image representation discussed above by collecting the AC DCT distributions over a hierarchy of sub-regions.

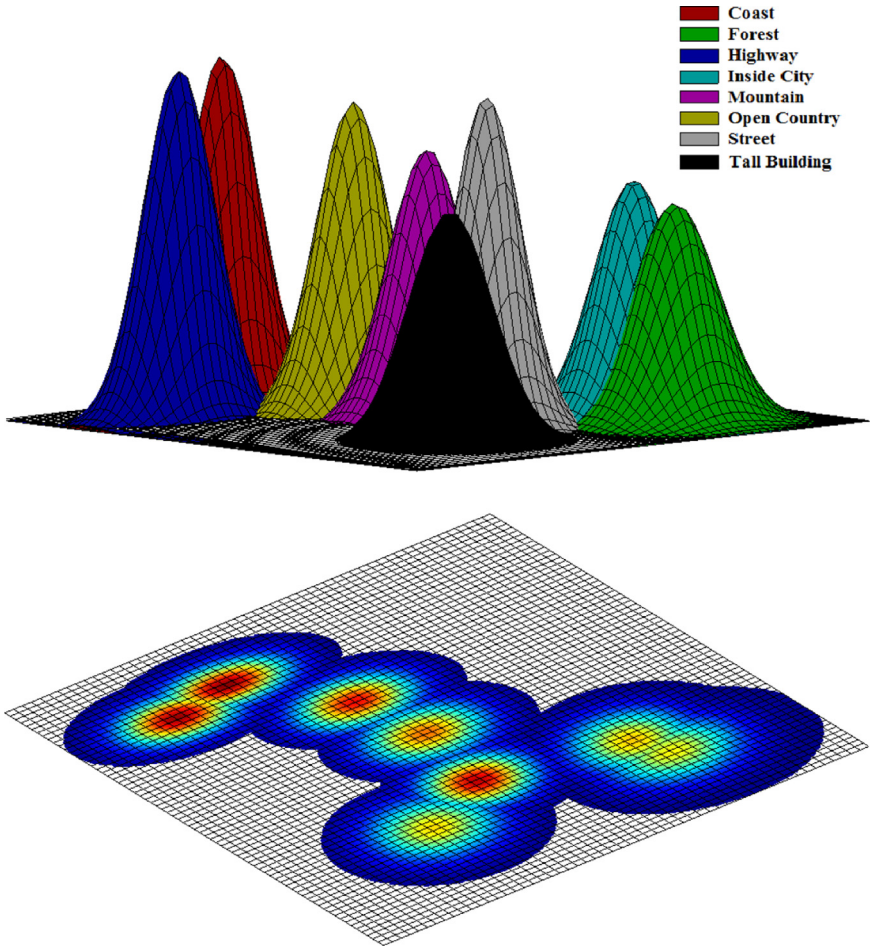

Fig. 4. 2-dimensional distributions (fitted with a Gaussian model) related to the Laplacian distribution parameters of the DCT frequency (0,1) and (1,0) in Fig. 1(b).

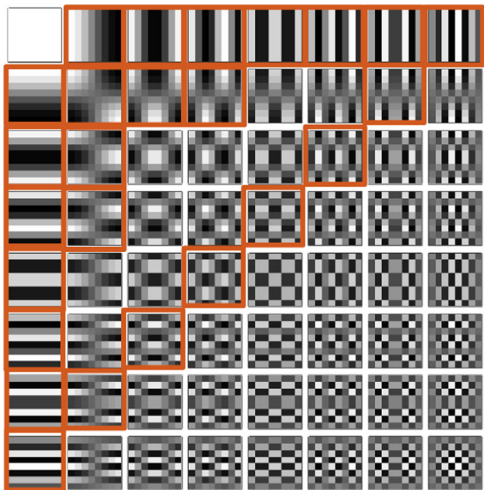

Fig. 5. Final AC DCT frequencies considered for representing the context of the scene (marked in red). (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

Specifically, the image is partitioned using three different modalities: horizontal, vertical and regular grid. These schemes are recursively applied to obtain a hierarchy of sub-regions as shown in Fig. 6. For each sub-region at each resolution level, the Laplacian parameters (μ and b) over the selected AC DCT coefficients are computed and concatenated to compose the feature vector, thus introducing spatial information. As in [17] we have used three levels in the hierarchy. The integral imaging approach [17,44] is exploited to efficiently compute the Laplacian parameters of the different AC DCT coefficients. The accuracy obtained on the aforementioned validation set, by considering the spatial hierarchy based representation was 0.8233%, improving the previous result of more than 6% (see Table 1).

Table 1
Accuracy of scene context classification on the validation dataset.

Approach	*Accuracy*
All frequencies (63 AC DCT basis)	0.7410
Selected AC DCT basis (Eq. (3))	0.7549
Selected AC DCT basis (Eq. (3)) and spatial hierarchy	0.8233

We can formalize the proposed DCT-GIST scene descriptor as following. Let $r^{l,s}$ be a sub-region of the image under consideration at level $l \in \{0,1,2\}$ of the subdivision scheme $s \in S = \{Horizontal, Vertical, Grid\}$ (see Fig. 6).[6] Let $H^{l,s}$ and $W^{l,s}$ be the number of 8×8 blocks of pixel with respect to the height and width of the region $r^{l,s}$. We indicate with the notation $B^{l,s}_{h,w,c}$, $h = 1,\ldots,H^{l,s}$, $w = 1,\ldots,W^{l,s}$, an 8×8 block of pixels of the region $r^{l,s}$ considering the color channel $c \in \{Y, C_b, C_r\}$. Let $D^{l,s}_{h,w,c}$ be the DCT components obtained from $B^{l,s}_{h,w,c}$ through a 2-dimensional DCT processing. We indicate with $D^{l,s}_{h,w,c}(i,j)$, $i = 1,\ldots,7$, $j = 1,\ldots,7$, the DCT components corresponding to the 2D DCT base (i,j) of Fig. 1(b). Let F be the set of the selected AC DCT basis defined above (Eq. (3)). Then, the

[6] Note that we define $r^{0,s}$ as the entire image under consideration for every $s \in S$.

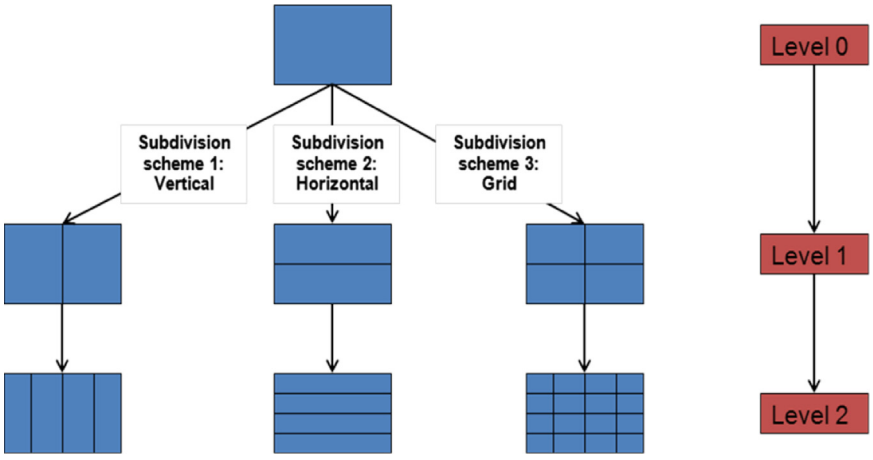

Fig. 6. Hierarchical subdivision of the image.

scene context descriptor of the region $r^{l,s}$ is computed as in the following equations (4)–(7):

$$\mu_c^{l,s}(0,0) = \frac{1}{H^{l,s}W^{l,s}} \sum_{h=1}^{H^{l,s}} \sum_{w=1}^{W^{l,s}} D_{h,w,c}^{l,s}(0,0) \tag{4}$$

$$b_c^{l,s}(0,0) = \frac{1}{H^{l,s}W^{l,s}} \sum_{h=1}^{H^{l,s}} \sum_{w=1}^{W^{l,s}} (D_{h,w,c}^{l,s}(0,0) - \mu_c^{l,s}(0,0))^2 \tag{5}$$

where Eq. (4) and (5) are evaluated for each $c \in \{Y, C_b, C_r\}$. The features in Eqs. (4) and (5) are related to the DC components of the DCT.

$$\mu_c^{l,s}(i,j) = Median(\{D_{h,w,c}^{l,s}(i,j) | h = 1,\ldots,H^{l,s}; w = 1,\ldots,W^{l,s}\}) \tag{6}$$

$$b_c^{l,s}(i,j) = \frac{1}{H^{l,s}W^{l,s}} \sum_{h=1}^{H^{l,s}} \sum_{w=1}^{W^{l,s}} \left| D_{h,w,c}^{l,s}(i,j) - \mu_c^{l,s}(i,j) \right| \tag{7}$$

where Eq. (6) and (7) are evaluated for each $(c,i,j) \in \{c \in \{Y, C_b, C_r\}; (i,j) \in F\}$. The features in Eqs. (6) and (7) are related to the 25 selected AC components of the DCT.

Let $[\boldsymbol{\mu}^{l,s}, \mathbf{b}^{l,s}]$ be the feature vector related to the region $r^{l,s}$ computed considering the Eqs. (4)–(7). The final image representation is obtained concatenating the representations $[\boldsymbol{\mu}^{l,s}, \mathbf{b}^{l,s}]$ of all the sub-regions in the spatial hierarchy (Fig. 6). The computational complexity to compute the proposed image representation is linear with respect to the number of 8×8 blocks composing the image region under consideration.

5. The image generation pipeline architecture

In this section we describe the system architecture to embed the scene context classification engine into an Image Generation Pipeline. The overall scheme is shown in Fig. 7. The "Scene Context Classification" module is connected to the "DCT" module. The "High resolution Pipe" block represents a group of algorithms devoted to the generation of high resolution images. This block is linked to the "Acquisition Information" block devoted to collect different information related to the image (e.g., exposure, gain, focus, white balance and). These information are used to capture and process the image itself. The "Viewfinder Pipe" block represents a group of algorithms which usually work on downscaled images to be shown in the viewfinder of a camera. The "Scene Context Classification" block works taking the input from the viewfinder pipe to determine the scene class of the image. The recognized class of the scene influences both the "Acquisition Information" and the "High resolution pipe" blocks in setting the parameters for the image acquisition. Moreover, the information obtained by the "Scene Context Classification" block can be exploited by the "Application Engine" block which can perform different operations according to the detected scene category. The "Memory lines" and "DMA" blocks provide the data arranged in 8×8 blocks to the "DCT" module for each image channel (Y, C_b, C_b). The "JPEG" block is the one that produces the final compressed image. The sub-blocks, composing the "Scene Context Classification" module, are described in the next subsections.

5.1. DCT coefficients accumulator

This block is directly linked to the "DCT" block, and thus it receives the DCT coefficients for the luminance and both chrominance channels. With reference to the hierarchical scheme shown in Fig. 6, this block accumulates DCT coefficients in histograms starting from the configuration having the smallest region size (e.g., level 2 of grid subdivision). For all the larger regions in the hierarchy, the computations can be performed by merging corresponding histogram bins previously computed at fine resolution level (e.g., the information already computed at level 2 can be exploited to compute the table at level 1 of grid subdivision).

5.2. Scene context representation

Starting from the histograms obtained by the "DCT Coefficients Accumulator" block, all the pairs of Laplacian parameters (μ and b) are computed by using the Laplacian fitting equations presented in Section 4. The scene context representation is then obtained by concatenating all the computed Laplacian parameters related to the selected DCT frequencies of all the sub-regions in the hierarchy for the three channels composing the image. In addition to this information, the mean and variance of the DC coefficients upon the hierarchy are computed exploiting the equations introduced in Section 4.

5.3. Classifier

The "Classifier" block takes the feature vector as input (i.e., the scene context representation) to perform the final scene context classification. It takes into account a classifier learned offline (i.e., the block "Model" in Fig. 7 which is learned out of the device). A Support Vector Machine is employed in our system architecture.

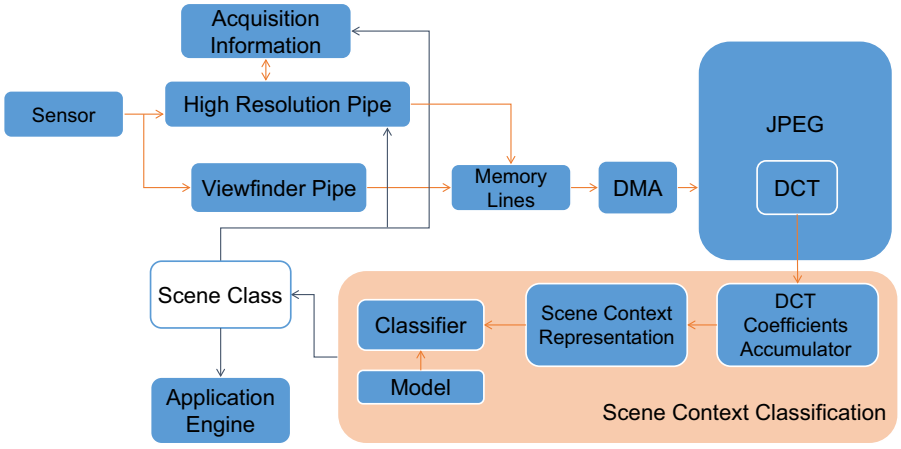

Fig. 7. Architecture of the IGP including the proposed scene context classification engine.

6. Experimental settings and results

In this section we report the experiments performed to quantitatively assess the effectiveness of the proposed DCT-GIST scene context descriptor with respect to other related approaches. In particular, we compare the performances obtained by the proposed representation model with respect to the ones achieved by the popular GIST descriptor [12] and the Roi+Gist Segmentation model proposed in [5]. Moreover, since the proposed representation is obtained collecting information on a spatial hierarchy, we have compared it with respect to the one which uses bags of textons on the same spatial hierarchy [17]. Finally, we describe how the architecture presented in Section 5 has been implemented on an IGP of a mobile device to demonstrate the effectiveness and the real-time performances of the proposed method. Experiments have been done by using a SVM and a 10-fold cross-validation protocol on each considered dataset. The images are first partitioned into 10 folds by making a random reshuffling of the dataset. Subsequently, 10 iterations of training and testing are performed such that within each iteration a different fold of the data is held-out for testing while the remaining folds are used for learning. The final results are obtained by averaging over the 10 runs. Since the proposed image representation can be used as input for any classifier, we reports also results obtained by exploiting the DCT-GIST representation with a Convolutional Neural Network classifier.

6.1. Proposed DCT-GIST representation vs GIST representation

To perform this comparison we have taken into account the scene dataset used in the paper introducing the GIST descriptor [12]. The dataset is composed by 2688 color images with resolution of 256×256 pixels (JPEG format) belonging to 8 scene categories: *Tall Building*, *Inside City*, *Street*, *Highway*, *Coast*, *Open Country*, *Mountain*, *Forest*. This dataset, together with the original code for computing the GIST descriptor are available on the web [45]. To better highlight the contribution of the different components involved in the proposed DCT-GIST representation (see Section 4) we have considered the following configurations (Table 2):

(A) Laplacian parameters of the 63 AC DCT components computed on Y channel;
(B) Laplacian parameters of the 25 selected AC DCT components computed on Y channel;
(C) Laplacian parameters of the 25 selected AC DCT components computed on Y channel and spatial hierarchy with 3 levels ($l=0,1,2$);

Table 2
The different configurations of the proposed DCT-GIST image representation.

DCT-GIST configuration	DCT frequencies	Image channels	Spatial hierarchy
(**A**)	All 63 AC components	Y	No
(**B**)	Selected 25 AC components	Y	No
(**C**)	Selected 25 AC components	Y	Yes
(**D**)	Selected 25 AC components + DC component	Y	Yes
(**E**)	Selected 25 AC components + DC component	YC_bC_r	No
(**F**)	Selected 25 AC components + DC component	YC_bC_r	Yes

(D) Laplacian parameters of the 25 selected AC DCT components computed on Y channel, mean and variance of the DC DCT components computed on Y channel, and spatial hierarchy with 3 levels ($l=0,1,2$);
(E) Laplacian parameters of the 25 selected AC DCT components computed on YC_bC_r channels, mean and variance of the DC DCT components computed on YC_bC_r channels;
(F) Laplacian parameters of the 25 selected AC DCT components computed on YC_bC_r channels, mean and variance of the DC DCT components computed on YC_bC_r channels, and spatial hierarchy with 3 levels ($l=0,1,2$).

Fig. 8 reports the average per class accuracy obtained considering all the above DCT-GIST representation configurations together with the results obtained employing the GIST descriptor [12]. The results show that the scene representation which considers only the Laplacian parameters of the 25 selected AC DCT frequencies fitted on the Y channel, i.e., the configuration (B), already obtains an accuracy of 75.20%. Encoding the information on the spatial hierarchy, i.e., configuration (C), is useful to improve the results of more than 6%. A small, but still useful, contribution is given by the color information obtained considering the DC DCT components, i.e., configuration (D). The proposed DCT-GIST representation obtains better results with respect to the GIST descriptor in both cases with and without spatial hierarchy (our with spatial hierarchy: 85.25%, our without spatial hierarchy: 84.60%, GIST: 84.28%). Table 3 reports the confusion matrix related to the proposed DCT-GIST representation corresponding to the configuration (F), whereas Table 4 shows the confusion matrix obtained by employing the GIST descriptor. One should not overlook that the proposed DCT-GIST representation has a very limited computational overhead for the image signature generation because

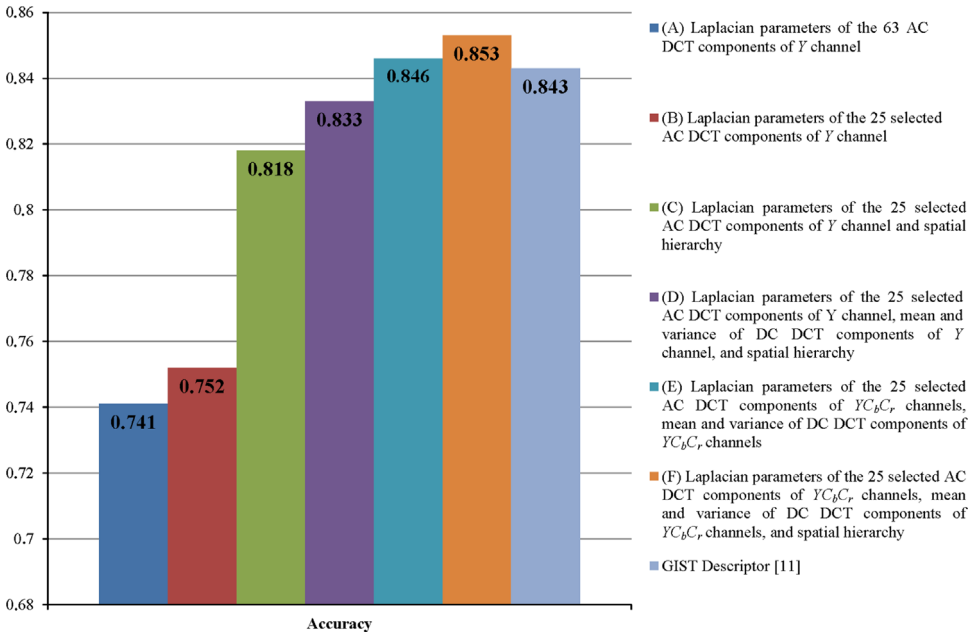

Fig. 8. Contribution of each component involved in the proposed DCT-GIST representation and comparison with respect to the GIST descriptor [12].

Table 3
Results obtained by exploiting the proposed DCT-GIST representation with configuration (F) on the 8 Scene Context Dataset [12]. Columns correspond to the inferred classes.

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**

it is directly computed by considering DCT coefficients already available from the JPEG encoder/format. Specifically, the computation of the image representation (F) requires about 1 operation per pixel (i.e., it is linear with respect to the image size). This highly reduces the complexity of the scene recognition system. Moreover, differently than GIST descriptor, the proposed representation is suitable for mobile platforms (e.g., smartphones and wearable cameras) since the DCT is already embedded in the Image Generation Pipeline, whereas the GIST descriptor needs extra overhead to compute the signature of the image and employs operations which are not present in the current IGP of single sensors imaging devices (e.g., FFT on the overall image). As detailed in the Sub-Section 6.5, the proposed DCT-GIST descriptor with configuration (F) can be computed in 15.9 ms on QVGA images (i.e., 320×240 pixels) with a 1 GHz Dual-core CPU. This computational time considers also the operations needed to compute the 8×8 DCT transformation of the input image. When the 8×8 DCT coefficients of the image are already available (e.g., in case of JPEG images or considering that these feature are computed into the IGP) the time needed to compute the proposed DCT-GIST descriptor is only 0.3 ms. As reported in [28] where an in-depth evaluation of the complexity of the GIST has been presented, the time needed to compute the GIST descriptor [12] on 64-bit 8-core computer and considering images of size 32×32 pixels is 35 ms. This means that the time needed to compute the proposed representation is at least half than the one needed to compute the GIST descriptor, and it is one order of magnitude less if the DCT coefficients are already available (i.e., JPEG format).

Further tests have been done to demonstrate the effectiveness of the proposed representation in discriminating the *Naturalness* and *Openness* of the scene [12]. Specifically, taking into account the definition given in [12], the *Naturalness* of the scene is related to the structure of a scene which strongly differs between man-made and natural environments. The notion of *Openness* is related to the open vs closed-enclosed environment, scenes with horizon vs no horizon, a vast or empty space vs a full, filled-in space [12]. A closed scene is a scene with small perceived depth, whereas an open scene is a scene with a big perceived depth. Information about *Naturalness* and/or *Openness* of the scene can be very useful in setting parameters of the algorithms involved in the image generation pipeline [13].

For the *Naturalness* experiment we have split the 8 scene dataset as in [12,15] by considering the classes *Coast*, *Open Country*, *Mountain* and *Forest* as *Natural* environments, whereas the classes *Tall Building*, *Inside City*, *Street* and *Highway* as belonging to the *Man-Made* environments. For the *Openness* experiment, the images belonging to the classes *Coast*, *Open Country*, *Street* and *Highway* have been considered as *Open* scenes, whereas the images of the classes *Forest*, *Mountain*, *Tall Building* and *Inside City* have been considered as *Closed* scenes. The results obtained employing the proposed representation (F) are reported in Tables 5 and 6. The obtained results closely match the performances of other state-of-the-art methods [15,17,26] by employing less computational resources.

Finally, we have considered the problem of recognizing four scene context usually available in the auto-scene mode of digital

Table 4
Results obtained exploiting the GIST representation [12] on the 8 Scene Context Dataset. Columns correspond to the inferred classes.

Confusion matrix	*Tall Building*	*Inside City*	*Street*	*Highway*	*Coast*	*Open Country*	*Mountain*	*Forest*
Tall Building	**0.83**	0.01	0.03	0.00	0.00	0.13	0.00	0.00
Inside City	0.00	**0.94**	0.00	0.00	0.05	0.01	0.00	0.01
Street	0.07	0.00	**0.82**	0.03	0.03	0.03	0.02	0.00
Highway	0.02	0.01	0.01	**0.84**	0.00	0.01	0.04	0.08
Coast	0.01	0.05	0.01	0.00	**0.86**	0.05	0.00	0.02
Open Country	0.14	0.04	0.02	0.00	0.05	**0.73**	0.01	0.00
Mountain	0.00	0.01	0.03	0.05	0.01	0.02	**0.87**	0.02
Forest	0.00	0.01	0.00	0.08	0.02	0.00	0.00	**0.88**

Table 5
Natural vs *Man-made* classification performances of the proposed DCT-GIST representation with configuration (F). Columns correspond to the inferred classes.

Confusion matrix	Natural	Man-made
Natural	**97.88**	2.12
Man-made	4.75	**95.25**

Table 6
Open vs *Closed* classification performances considering the proposed DCT-GIST representation with configuration (F). Columns correspond to the inferred classes.

Confusion matrix	Open	Closed
Open	**94.17**	5.83
Closed	4.63	**95.37**

Table 7
Results obtained by the proposed DCT-GIST representation with configuration (F) on four classes usually used in the auto-scene mode of consumer digital cameras.

Confusion matrix	Landscape	Man-made outdoor	Portrait	Snow
Landscape	**87.76**	1.22	0.61	10.41
Man-made outdoor	3.78	**91.33**	2.22	2.67
Portrait	1.02	1.84	**94.29**	2.86
Snow	9.62	1.13	3.02	**86.23**

Table 8
Results obtained by GIST [12] on four classes usually used in the auto-scene mode of consumer digital cameras.

Confusion matrix	Landscape	Man-made outdoor	Portrait	Snow
Landscape	**84.69**	3.27	0.20	11.84
Man-made outdoor	4.44	**87.78**	2.44	5.33
Portrait	0.41	3.47	**91.84**	4.29
Snow	11.70	3.40	4.34	**80.57**

consumer cameras: *Landscape*, *Man-Made Outdoor*, *Portrait*, *Snow*. To this purpose we have collected 2000 color images (i.e., 500 per class) with resolution 640×480 pixels from Flickr . This dataset has been used to perform a comparative test of the proposed DCT-GIST image representation with configuration (F) with respect to the popular GIST descriptor. The results are reported on Tables 7 and 8. The proposed image representation obtained an average accuracy of 89.80%, whereas GIST achieved 86.07%.

Table 9
Results obtained on the 15 Scene Dataset [18].

Bags of textons with spatial hierarchy [17]	79.43%
Proposed DCT-GIST representation with configuration (D)	78.45%
GIST representation [12]	73.25%

6.2. Proposed representation vs bags of textons on spatial hierarchy

Since the proposed scene context representation works exploiting information collected on spatial hierarchy, we have compared it with respect to the method presented in [17], where Bags of Textons are collected for each region in the spatial hierarchy to represent the images for scene classification purposes. For this comparison we have considered the 15 Scene Classes Dataset used in [18]. This dataset is an augmented version of the 8 Scene Classes Dataset [12]. The dataset is composed by 4485 images of the following fifteen categories: *highway*, *inside of cities*, *tall buildings*, *streets*, *forest*, *coast*, *mountain*, *open country*, *suburb residence*, *bedroom*, *kitchen*, *living room*, *office*, *industrial* and *store*. Since a subset of the images of the dataset does not have color information, the tests on the 15 Scene Classes Dataset have been performed taking into account only the *Y* channel and using the DCT-GIST scene descriptor with configuration (D) (see Table 2). The results obtained on this dataset are reported in Table 9. The average per class accuracy achieved by the proposed approach is 78.45%, whereas the method which exploit textons distributions on spatial hierarchy [17] obtained an accuracy of 79.43%. Both representations outperform the GIST one, which obtains 73.25% of accuracy on this dataset. Although the results are slightly in favor for the method proposed in [17] (of less than 1%), one should not forget that the proposed DCT-GIST representation is suitable for an implementation on the image generation pipeline of single sensor devices, whereas the method in [17] requires extra memory to store Textons vocabularies (i.e., hardware costs for industry) as well as a bigger computational overhead to represent the image to be classified (e.g., convolution with bank of filters, computation of the Textons distributions for every sub-regions, etc.). Specifically, considering an image stored in JPEG format, the computation of the Bag of Textons signature in [17] requires the convolution of the image with a bank of 24 filters of size 49×49 (i.e., $49 \times 49 \times 24$ operations per pixel), and the computation of the similarity of each pixel responses with respect to the Textons vocabulary (i.e., T operations per pixels, where T is the number of Textons in the vocabulary). Hence, the computational time needed to build the Bag of Textons signature is much higher than the one to compute the proposed DCT-GIST representation (i.e., linear with respect to the number of 8×8 blocks composing the image region under consideration).

We have performed one more test to assess the ability of the proposed representation in discriminating among *Indoor* vs *Outdoor* scenes. This prior can be very useful for autofocus, auto-exposure and white balance algorithms. To this aim we have

Table 10
Indoor vs *outdoor* classification performances considering the proposed DCT-GIST representation with configuration (D). Columns correspond to the inferred classes.

Confusion matrix	Indoor	Outdoor
Indoor	**89.75**	10.25
Outdoor	3.86	**96.14**

divided the images of the 15 Scene Classes Dataset as indoor vs outdoor images. The classification results are reported in Table 10. Again the results confirm that the proposed representation can be employed to distinguish classes of scenes at superordinate level of description [12].

6.3. DCT-GIST evaluation on the MIT-67 indoor scene dataset

To further assess the proposed image representation we have performed tests by considering the challenging problem of discriminating among different indoor scenes categories. To this aim we have considered the MIT-67 Indoor Scene Dataset [5] which contains 67 Indoor categories and a total of 15,620 images. The MIT-67 dataset is one of the largest dataset of scenes available so far. In performing the experiments we have considered the testing protocol used in [5] (i.e., same training and testing images). The tests have been done considering the configuration (F) of the proposed representation (see Table 2). The proposed descriptor has been compared with respect to the GIST [12] as well as with respect to the model called ROI+Gist Segmentation (RGS) which has been introduced in the paper related to MIT-67 dataset [5]. The RGS representation model combine both global (i.e., GIST) and local information (i.e., spatial pyramid of visual words on ROIs obtained by segmenting the image). Hence the RGS is able to take into account global spatial properties of the scenes and the concepts/objects they contain.

The experiments pointed out that our DCT-GIST scene descriptor achieves an average per-class accuracy of 26.7%, which is greater than the one obtained by both GIST (less than 22%) and RGS model (25.05%) [5]. Table 11 reports the per-class accuracies obtained with both the proposed DCT-GIST and the RGS model. Also in this case the DCT-GIST descriptor obtains comparable recognition performances with respect to the state-of-the-art, and outperforms the state-of-the-art in terms of computational complexity (i.e., RGS model needs to compute the GIST with its related computational complexity, needs a segmentation step, and also uses a spatial based bag of visual word model. Hence, DCT-GIST is more suitable for the Imaging Generation Pipeline in terms of both time and memory resources).

6.4. Instant scene context classification on mobile device

The experiments presented in Sections 6.1–6.3 have been performed on representative datasets used as benchmark in the literature. For those tests the DCT-GIST scene context representation has been obtained directly by extracting the DCT information from the compressed domain (JPEG format). The main contribution of this paper is related to the possibility to obtain a signature for the scene context directly into the image generation pipeline of a mobile platform, taking into account the architecture presented in Section 5. To this aim we have implemented the proposed architecture on a Nokia N900 smartphone [46]. This mobile platform has been chosen because it has less computational power of the other smartphones (i.e., the scene context classification engine should able to classify in real-time independently of the computational power of the device). Moreover, with the chosen mobile platform, the FCam API can be employed to work within the Image Generation Pipeline of the device [47,48]. This allows to effectively build the proposed architecture and test it with real settings. Although the limited resources of the hand-held device, the

Table 11
Recognition results of the proposed DCT-GIST descriptor on the MIT-67 dataset [5]. The proposed representation is compared with respect to the ROI+Gist Segmentation model [5].

Classes	Proposed DCT-GIST	RGS model [5]
elevator	**71.40**	61.90
greenhouse	**65.00**	50.00
concert hall	**60.00**	45.00
inside bus	**56.50**	39.10
corridor	**52.40**	38.10
bowling	**50.00**	45.00
buffet	50.00	**55.00**
classroom	**50.00**	**50.00**
cloister	**50.00**	45.00
casino	**47.40**	21.10
hospital room	**45.00**	35.00
pantry	**45.00**	25.00
auditorium	44.40	**55.60**
church inside	42.10	**63.20**
library	**40.00**	**40.00**
bathroom	**38.90**	33.30
clothing store	**38.90**	22.20
tv studio	**38.90**	27.80
children room	**33.30**	5.60
closet	33.30	**38.90**
inside subway	**33.30**	23.80
florist	31.60	**36.80**
studio music	31.60	**36.80**
airport inside	**30.00**	10.00
kinder garden	**30.00**	5.00
movie theater	**30.00**	15.00
dental office	28.60	**42.90**
grocery store	28.60	**38.10**
dining room	**27.80**	16.70
meeting room	**27.30**	9.10
video store	**27.30**	**27.30**
art studio	**25.00**	10.00
living room	**25.00**	15.00
lobby	**25.00**	10.00
nursery	25.00	**35.00**
prison cell	**25.00**	10.00
restaurant	**25.00**	5.00
computer room	22.20	**44.40**
garage	22.20	**27.80**
bakery	**21.10**	15.80
game room	20.00	**25.00**
stairscase	20.00	**30.00**
train station	20.00	**35.00**
subway	**19.00**	9.50
bar	16.70	**22.20**
gym	16.70	**27.80**
deli	15.80	**21.10**
bedroom	**14.30**	**14.30**
kitchen	14.30	**23.80**
locker room	14.30	**38.10**
laundromat	13.60	**31.80**
toystore	**13.60**	**13.60**
restaurant kitchen	**13.00**	4.30
fast-food restaurant	11.80	**23.50**
mall	**10.00**	0.00
hair salon	**9.50**	**9.50**
office	**9.50**	0.00
warehouse	**9.50**	**9.50**
laboratory wet	**9.10**	0.00
operating room	5.30	**10.50**
bookstore	5.00	**20.00**
pool inside	5.00	**25.00**
jewellery shop	**4.50**	0.00
museum	**4.30**	**4.30**
shoe shop	0.00	**5.30**
waiting room	0.00	**19.00**
wine cellar	0.00	**23.80**
Average	**26.70**	25.05

Fig. 9. Example scene context classification of the system implemented on the Nokia N900.

Table 12
Comparison of DCT-GIST with respect to GIST [12] employing Convolutional Neural Network classifier on the 8 Scene Dataset.

Proposed DCT-GIST	GIST
86.49	86.47

Table 13
Comparison of DCT-GIST with respect to GIST [12] employing Convolutional Neural Network classifier on the MIT-67 Dataset.

Proposed DCT-GIST	GIST
28.81	22.68

implemented system works in real-time as demonstrated by the video available at the following URL: http://iplab.dmi.unict.it/DCT-GIST.

For the implemented system we have used a SVM model learned offline on the 8 Scene Context Dataset (see Section 6.1) and the configuration (F) for the DCT-GIST representation (see Table 2). The scene context representation is computed on the fly during the generation of the image to be displayed in the view-finder. The implemented architecture can also perform classification of images already stored in the mobile (Fig. 9).

The proposed DCT-GIST based scene context classifier has been also tested on a NovaThor U9500 with Android OS. The board mounts a 1 GHz Dual-core ARM Cortex-A9 CPU. The computational time performances have been evaluated by considering the average latencies of the different scene classification blocks on a set of QVGA images. We have measured the computational time of all the steps involved in the scene classification engine: DCT computation, DCT-GIST image representation with configuration (F) (see Table 2) and the SVM classification. The DCT computation required 15.6 ms on the average (this value could be disregarded when DCT coefficients are directly provided by the integrated JPEG encoder or by working directly on compressed domain). The overall computational time to build the image signature with configuration (F) (i.e., the one with spatial hierarchy and all the three image channels of the image) was only 0.3 ms. Finally, the SVM classification required 117.4 ms. This test confirmed that the proposed image signature can be computed in realtime within a mobile platform. Note that the GIST descriptor [12] is not suitable for the IGP (i.e., FFT is not present into the IGP) and it is known from [28] that the time needed for its computation is ≥ 35 ms (i.e., higher than the one needed to compute proposed DCT-GIST descriptor).

6.5. Further experiments exploiting convolutional neural network classifier

The proposed DCT-GIST representation can be used with any classifier. The test reported so far have been performed by employing the SVM classifier to compare our approach with respect to the other compared scene descriptors [12,17,5]. To further test the proposed DCT-GIST representation with respect to the GIST we have employed Convolutional Neural Network as classifier. The results of this comparison are reported in Tables 12 and 13. Note that the average per class accuracy is in favor of the proposed descriptor. The results obtained with CNN are slightly better than the one obtained with SVM in almost all cases.

7. Conclusion and future works

This paper introduces the DCT-GIST image representation to be exploited for scene context classification on mobile platforms. The proposed scene descriptor is based on the statistics of the DCT coefficients. Starting from the knowledge that the distribution of the AC DCT coefficients can be approximated by Laplacian distributions, and from the observation that different scene context present differences in the Laplacian scales, we proposed a signature of the scene that can be efficiently computed directly in the compressed domain (from JPEG format), as well as in the image generation pipeline of single sensor devices (e.g., smartphones, consumer digital cameras, and wearable smart cameras). The effectiveness of the proposed scene context descriptor has been demonstrated on representative datasets by comparing it with respect to the popular GIST descriptor [12] and the representation based on textons distributions on spatial hierarchy [17] and the ROI+Gist segmentation model [5]. Moreover, the proposed scene context recognition architecture has been implemented and tested on a real acquisition pipeline of a mobile phone to demonstrate the real-time performances of the overall system. Differently than other state-of-the-art scene descriptors, the computation of the proposed signature does not need extra information to be stored in memory (e.g., visual vocabulary) or complex operations (e.g., convolutions, FFT, learning phase). The proposed holistic scene representation provides an efficient way to obtain information about the context of the scene which can be extremely useful as first step for object detection and context driven focus attention algorithms by priming typical objects, scales and locations [9,10]. It can be also exploited to have priors for setting the parameters of the algorithm involved in the IGP (e.g., white balance) to improve the quality of the final acquired image [13]. Future works could consider the integration of camera metadata related to the image capture conditions to improve recognition accuracy [49,50].

Conflict of interest

None declared.

Acknowledgments

The authors would like to thank Nokia Research Center [51] for providing the N900 smartphones. This research has been supported by STMicroelectronics [52].

References

[1] D. Marr, Vision: A Computational Investigation into the Human Representation and Processing of Visual Information, W.H. Freeman, Henry Holt and Co., Inc., New York, NY, USA, 1982.
[2] I. Biederman, Aspects and extension of a theory of human image understanding, in: Computational Processes in Human Vision: An Interdisciplinary Perspective, 1988.
[3] A. Oliva, A. Torralba, Building the gist of a scene: the role of global image features in recognition, Vis. Percept.: Prog. Brain Res. 155 (2006) 251–256.
[4] J. Vogel, A. Schwaninger, C. Wallraven, H.H. Bülthoff, Categorization of natural scenes: local versus global information and the role of color, ACM Trans. Appl. Percept. 4 (3) (2007) 1–21.
[5] A. Quattoni, A. Torralba, Recognizing indoor scenes, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009.
[6] A. Torralba, K.P. Murphy, W.T. Freeman, M.A. Rubin, Context-based vision system for place and object recognition, in: IEEE International Conference on Computer Vision (ICCV-03), 2003, pp. 273–280.
[7] P. Ladret, A. Guérin-Dugué, Categorisation and retrieval of scene photographs from JPEG compressed database, Pattern Anal. Appl. 4 (2001) 185–199.
[8] J. Vogel, B. Schiele, Semantic modeling of natural scenes for content-based image retrieval, Int. J. Comput. Vis. 72 (2) (2007) 133–157.
[9] A. Torralba, Contextual priming for object detection, Int. J. Comput. Vis. 53 (2) (2003) 169–191.
[10] A. Torralba, S. Pawan, Statistical context priming for object detection, in: IEEE International Conference on Computer Vision, 2001.
[11] A. Torralba, A. Oliva, Depth estimation from image structure, IEEE Trans. Pattern Anal. Mach. Intell. 24 (9) (2002) 1226–1238.
[12] A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope, Int. J. Comput. Vis. 42 (2001) 145–175.
[13] S. Bianco, G. Ciocca, C. Cusano, R. Schettini, Improving color constancy using indoor–outdoor image classification, IEEE Trans. Image Process. 17 (12) (2008) 2381–2392.
[14] S. Battiato, A.R. Bruna, G. Messina, G. Puglisi, Image Processing for Embedded Devices, Bentham Science Publisher, Bentham Science Publishers, The Netherlands, 2010.
[15] G.M. Farinella, S. Battiato, Scene classification in compressed and constrained domain, IET Comput. Vis. 5 (5) (2011) 320–334.
[16] E.Y. Lam, J.W. Goodman, A mathematical analysis of the DCT coefficient distributions for images, IEEE Trans. Image Process. 9 (10) (2000) 1661–1666.
[17] S. Battiato, G.M. Farinella, G. Gallo, D. Ravì, Exploiting textons distributions on spatial hierarchy for scene classification, Eurasip J. Image Video Process. (2010) 1–13.
[18] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR-06), 2006, pp. 2169–2178.
[19] L.W. Renninger, J. Malik, When is scene recognition just texture recognition? Vis. Res. 44 (2004) 2301–2311.
[20] A. Bosch, A. Zisserman, X. Muñoz, Scene classification via PLSA, in: European Conference on Computer Vision (ECCV-06), 2006, pp. 517–530.
[21] S. Battiato, G.M. Farinella, G. Gallo, D. Ravì, Scene categorization using bag of textons on spatial hierarchy, in: IEEE International Conference on Image Processing, 2008, pp. 2536–2539.
[22] G. Csurka, C. Dance, L. Fan, J. Willamowski, C. Bray, Visual categorization with bags of keypoints, in: ECCV International Workshop on Statistical Learning in Computer Vision, 2004, pp. 1–22.
[23] S. Battiato, G.M. Farinella, G. Gallo, D. Ravì, Spatial hierarchy of textons distributions for scene classification, in: International Conference on Multi-Media Modeling, Lecture Notes in Computer Science, vol. 5371, 2009, pp. 333–343.
[24] A. Bosch, X. Munoz, R. Martí, Review: which is the best way to organize/classify images by content? Image Vis. Comput. 25 (6) (2007) 778–791.
[25] A. Torralba, A. Oliva, Statistics of natural image categories, Netw.: Comput. Neural Syst. 14 (2003) 391–412.
[26] A. Torralba, A. Oliva, Semantic organization of scenes using discriminant structural templates, in: IEEE International Conference on Computer Vision (ICCV-99), 1999, pp. 1253–1258.
[27] H. Grabner, F. Nater, M. Druey, L.V. Gool, Visual interestingness in image sequences, in: ACM International Conference on Multimedia, 2013.
[28] M. Douze, H. Jégou, H. Sandhawalia, L. Amsaleg, C. Schmid, Evaluation of GIST descriptors for web-scale image search, in: ACM International Conference on Image and Video Retrieval, 2009, pp. 1–8.
[29] Z. Lu, K. Grauman, Story-driven summarization for egocentric video, in: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2714–2721.
[30] J. Luo, M.R. Boutell, Natural scene classification using overcomplete ICA, Pattern Recognit. 38 (10) (2005) 1507–1519.
[31] G. Farinella, S. Battiato, G. Gallo, R. Cipolla, Natural versus artificial scene classification by ordering discrete Fourier power spectra, in: Structural, Syntactic, and Statistical Pattern Recognition, Lecture Notes in Computer Science, vol. 5342, Springer, Berlin, Heidelberg, 2008, pp. 137–146.
[32] G.K. Wallace, The JPEG still picture compression standard, Commun. ACM 34 (4) (1991) 18–34.
[33] W.K. Pratt, Digital Image Processing, John Wiley & Sons, Inc., New York, NY, USA, 1978.
[34] J.D. Eggerton, Statistical distributions of image DCT coefficients, Comput. Electr. Eng. 12 (1986) 137–145.
[35] F. Müller, Distribution shape of two-dimensional DCT coefficients of natural images, Electron. Lett. 29 (1993) 1935–1936.
[36] T. Eude, R. Grisel, H. Cherifi, R. Debrie, On the distribution of the DCT coefficients, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 5, 1994, pp. 365–368.
[37] S. Smoot, L.A. Rowe, Study of DCT coefficient distributions, in: SPIE Symposium on Electronic Imaging, vol. 2657, 1996, pp. 403–411.
[38] G.S. Yovanof, S. Liu, Statistical analysis of the DCT coefficients and their quantization, in: Conference Record of the Thirtieth Asilomar Conference on Signals, Systems and Computers, 1996.
[39] C.-C. Chang, J.-C. Chuang, Y.-S. Hu, Retrieving digital images from a JPEG compressed image database, Image Vis. Comput. 22 (6) (2004) 471–484.
[40] G. Schaefer, Content-based image retrieval: advanced topics, in: T. Czachórski, S. Kozielski, U. Stańczyk (Eds.), Man-Machine Interactions 2, Advances in Intelligent and Soft Computing, vol. 103, Springer, Berlin Heidelberg, 2011, pp. 31–37.
[41] E. Lam, Analysis of the DCT coefficient distributions for document coding, IEEE Signal Process. Lett. 11 (2) (2004) 97–100.
[42] R.M. Norton, The double exponential distribution: using calculus to find a maximum likelihood estimator, Am. Stat. 38 (2) (1984) 135–136.
[43] G. Yu, G. Sapiro, S. Mallat, Image modeling and enhancement via structured sparse model selection, in: IEEE International Conference on Image Processing (ICIP), 2010, pp. 1641–1644.
[44] P. Viola, M.J. Jones, Robust real-time face detection, Int. J. Comput. Vis. 57 (2) (2004) 137–154.
[45] A. Oliva, A. Torralba, Gist Descriptor, 2001, URL ⟨http://people.csail.mit.edu/torralba/code/spatialenvelope/⟩.
[46] S. Battiato, G. Farinella, M. Guarnera, D. Ravì, V. Tomaselli, Instant scene recognition on mobile platform, in: European Conference on Computer Vision (ECCV) – Workshops and Demonstrations, Lecture Notes in Computer Science, vol. 7585, 2012, pp. 655–658.
[47] A. Adams, E.-V. Talvala, S.H. Park, D.E. Jacobs, B. Ajdin, N. Gelfand, J. Dolson, D. Vaquero, J. Baek, M. Tico, H.P.A. Lensch, W. Matusik, K. Pulli, M. Horowitz, M. Levoy, The frankencamera: an experimental platform for computational photography, ACM Trans. Gr. 29 (4) (2010) 1–12.
[48] F. Garage, Fcam api, ⟨http://fcam.garage.maemo.org/⟩, 2012.
[49] M. Boutell, J. Luo, Bayesian fusion of camera metadata cues in semantic scene classification, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004, pp. 623–630.
[50] M. Boutell, J. Luo, Beyond pixels: exploiting camera metadata for photo classification, Pattern Recognit. 38 (6) (2005) 935–946.
[51] Nokia Research Center, Visual Computing and Ubiquitous Imaging, 2013, URL ⟨http://research.nokia.com/⟩.
[52] STMicroelectronics, Advanced System Technology – Computer Vision Group, URL ⟨http://www.st.com/⟩.

Giovanni Maria Farinella received the M.S. degree in computer science (egregia cum laude) from the University of Catania, Italy, in 2004, and the Ph.D. degree in computer science in 2008. He joined the Image Processing Laboratory (IPLAB) at the Department of Mathematics and Computer Science, University of Catania, in 2008, as a Contract Researcher. He is a Contract Professor of Computer Vision at the Academy of Arts of Catania (since 2004) and Adjunct Professor of Computer Science at the University of Catania (since 2008). His research interests lie in the fields of computer vision, pattern recognition and machine learning. He has edited four volumes and coauthored more than 60 papers in international journals, conference proceedings and book chapters. He is a co-inventor of four international patents. He serves as a reviewer and on the programme committee for major international journals and international conferences. He founded (in 2006) and currently directs the International Computer Vision Summer School.

Daniele Ravì was born in Sant'Agata di Militello (ME), Italy, in 1983. He received the Master Degree in Computer Science (summa cum laude) in 2007 from University of Catania. From 2008 to 2010 he worked at STMicroelectronics (Advanced System Technology Imaging Group) as consultant. He recently has finished his Ph.D. in Computer Science and now he is a research engineer at Visual Atoms. His interests lie in the fields of computer vision, image analysis, visual search and machine learning.

Valeria Tomaselli is a SW Design Senior Engineer and Project Leader at STMicroelectronics in Catania. She received the Master Degree in Software Engineering (summa cum laude) in 2003 from University of Catania. From 2003 she has been working at STMicroelectronics, in the Advanced System Technology group, where she researches innovative algorithms for Digital Still Camera and Mobile Imaging applications in the image processing and computer vision fields. She is author of patents and papers about image processing and computer vision, and she also serves as a reviewer. She has been also involved in many national and international research projects.

Mirko Guarnera received his Master Degree in Electronic Engineering from the University of Palermo and the Ph.D. from University of Messina. He joined STMicroelectronics at the AST Labs in Catania in 1999, where he currently holds the position of R&D Project Manager. He is IEEE member and member of the technical committee of SPIE Electronic Imaging – Digital Photography conference. His research interests include image processing and pattern recognition for camera, TV, printers and projectors. He is author of many Papers in journals, book chapters and Patents.

Sebastiano Battiato received his degree in computer science (summa cum laude) in 1995 from University of Catania and his Ph.D. in computer science and applied mathematics from University of Naples in 1999. From 1999 to 2003 he was the leader of the "Imaging" team at STMicroelectronics in Catania. He joined the Department of Mathematics and Computer Science at the University of Catania as assistant professor in 2004 and became associate professor in the same department in 2011. His research interests include image enhancement and processing, image coding, camera imaging technology and multimedia forensics. He has edited 4 books and co-authored more than 150 papers in international journals, conference proceedings and book chapters. He is a co-inventor of about 15 international patents, reviewer for several international journals, and he has been regularly a member of numerous international conference committees. Prof. Battiato has participated in many international and national research projects. Chair of several international events (IWCV2012, ECCV2012, VISAPP 2012–2013–2014, ICIAP 2011, ACM MiFor 2010–2011, SPIE EI Digital Photography 2011–2012–2013, etc.). He is an associate editor of the IEEE Transactions on Circuits and System for Video Technology and of the SPIE Journal of Electronic Imaging. Guest editor of the following special issues: "Emerging Methods for Color Image and Video Quality Enhancement" published on EURASIP Journal on Image and Video Processing (2010) and "Multimedia in Forensics, Security and Intelligence" published on IEEE Multimedia Magazine (2012). He is the recipient of the 2011 Best Associate Editor Award of the IEEE Transactions on Circuits and Systems for Video Technology. He is director (and co-founder) of the International Computer Vision Summer School (ICVSS), Sicily, Italy. He is a senior member of the IEEE.