www.ietdl.org

Published in IET Computer Vision
Received on 3rd May 2010
Revised on 17th January 2011
doi: 10.1049/iet-cvi.2010.0056

ISSN 1751-9632

Scene classification in compressed and constrained domain

G.M. Farinella S. Battiato

Image Processing Laboratory, Dipartimento di Matematica e Informatica, Università di Catania, Catania, Italy
E-mail: gfarinella@dmi.unict.it

Abstract: Holistic representations of natural scenes are an effective and powerful source of information for semantic classification and analysis of images. Despite the technological hardware and software advances, consumer single-sensor imaging devices technology are quite far from the ability of recognising scenes and/or to exploit the visual content during (or after) acquisition time. The frequency domain has been successfully exploited to holistically encode the content of natural scenes in order to obtain a robust representation for scene classification. The authors exploit a holistic representation of the scene in the discrete cosine transform domain fully compatible with the JPEG format. The advised representation is coupled with a logistic classifier to perform classification of the scene at superordinate level of description (e.g. natural against artificial), or to discriminate between multiple classes of scenes usually acquired by a consumer imaging device (e.g. portrait, landscape and document). The proposed method is able to work in constrained domain. Experiments confirm the effectiveness of the proposed method. The obtained results closely match state-of-the-art methods in terms of accuracy outperforming in terms of computational resources.

1 Introduction and motivations

Before shooting a scene, a photographer adjusts 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 could be 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). Typical consumer devices acquire digital images by using digital sensors (e.g. CCD/CMOS). Quality improvement is obtained by increasing the resolution of the sensor or by using *ad hoc* image processing algorithms [1, 2]. Acquisition of colour images requires the presence of different sensors for different colour channels. Manufacturers reduce the cost and complexity by placing a colour filter array (CFA) on top of a single sensor, which is basically a monochromatic device, to acquire colour information of the true visual scene. The overall performances of any device are the result of a mixture of different components including hardware and software capabilities and, not ultimately, overall design (i.e. shape, weight, style, etc.). Typical imaging pipelines implemented in single-sensor cameras are designed to find a trade-off between sub-optimal solutions (devoted to solve imaging acquisition) and technological problems (e.g. colour balancing, thermal noise, etc.) in the context of limited hardware resources. State-of-the-art techniques to process multichannel pictures, obtained through peculiar processing of CFA images, include demosaicing, enhancement, denoising, compression and also *ad hoc* matrixing and colour balancing techniques devoted to preprocess input data coming from the sensor. The overall image generation pipeline (IGP) is aimed to reconstruct the final image exploiting all the information acquired by sensor to achieve the 'best' possible image.

Despite the technological hardware and software advances [3–9], consumer single-sensor imaging devices technology (see [1, 2] for a recent review in the field) are quite far from the ability of recognising scenes and/or to exploit the visual content during (or after) acquisition time. Although there is progress in the area of scene understanding into post-acquisition time, a framework describing this task during (or right after) acquisition time is still missing in the literature. The need for the development of solution for scene recognition systems to be embedded in consumer imaging devices domain, where limited resources are available, is confirmed by the growing interest of consumer devices industry [10].

We propose a scene categorisation engine in which the holistic representation of the scene is built exploiting features extracted on discrete cosine transform (DCT) domain. A logistic classifier is trained and then used to infer the category of a new observed scene. The proposed approach is fully compatible with JPEG format and may be easily employed on constrained domain.

Images are represented as histograms of oriented blocks [11] coupled with statistical weights to evaluate how important is an orientation in discriminating the classes under consideration [12]. Two local features are extracted and used to represent each 8×8 spatial block belonging to a considered image (Fig. 1): the dominant orientation of the

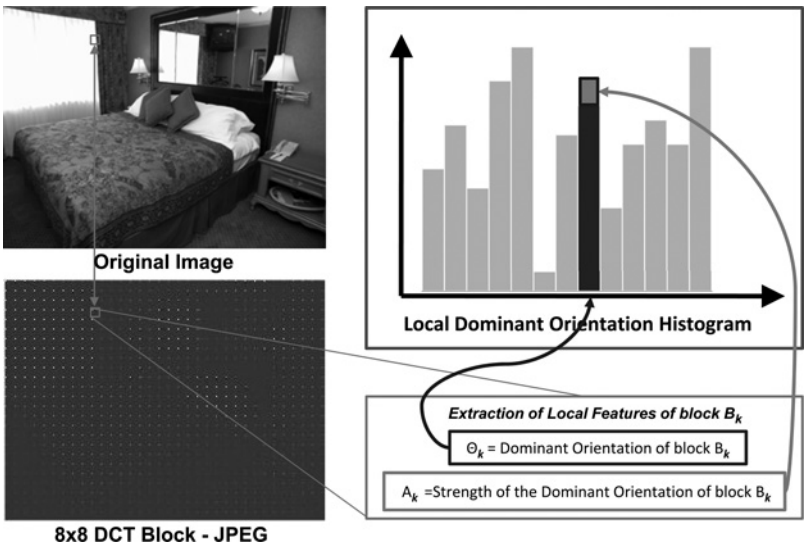

Fig. 1 *Image is represented as histogram of LDOs weighted by their strengths*

block and the strength of the dominant orientation. These two local information are extracted, for each 8 × 8 spatial block within the image under consideration, directly on compressed domain considering the corresponding 8 × 8 DCT block [13]. Specifically, the ratio between the sum of the DCT coefficients corresponding to the horizontal frequencies and the sum of DCT coefficients corresponding to the vertical frequencies is used to establish the tangent of the local dominant orientation (LDO) angle of an 8 × 8 spatial block (Fig. 2). The overall AC energy of each block is related to the strength of the LDO of an 8 × 8 spatial block (Fig. 3). The extracted local features are then used to build a holistic representation of the image as a distribution of LDO. This representation is coupled with TF-IDF weighting scheme [12] to statistically capture the most discriminative orientations between classes of scenes.

In the context of single-sensor imaging devices, the primary contributions of this work can be summarised as follows:

- Despite the proposed approach works on compressed and constrained domain, the classification rate of the proposed approach closely matches the other state-of-the-art methods.
- A simple probabilistic model is used to perform classification. The classifier is trained offline just considering a data set properly collected. The learned parameters of the model may be used online to classify a new observed scene through a simple decision function. A very compact low-dimensional vector of parameters is maintained to perform classification.
- It is directly implemented in the DCT domain and compliant with the JPEG format. Local features are picked-out by using simple operations in compressed DCT domain. Bank of filters are not used during features extraction phase. For embedded imaging devices, the DCT data can be obtained without additional effort just at the end of the IGP; alternatively such information can be used just before JPEG quantisation phase.
- The proposed approach is able to work on the thumbnail version of the acquired images.
- The global representation of the scene is obtained grouping together the extracted local information in a very compact low-dimensional vector; no extra information (e.g. visual vocabulary) need to be stored in memory to build and manage the holistic representation of a scene.

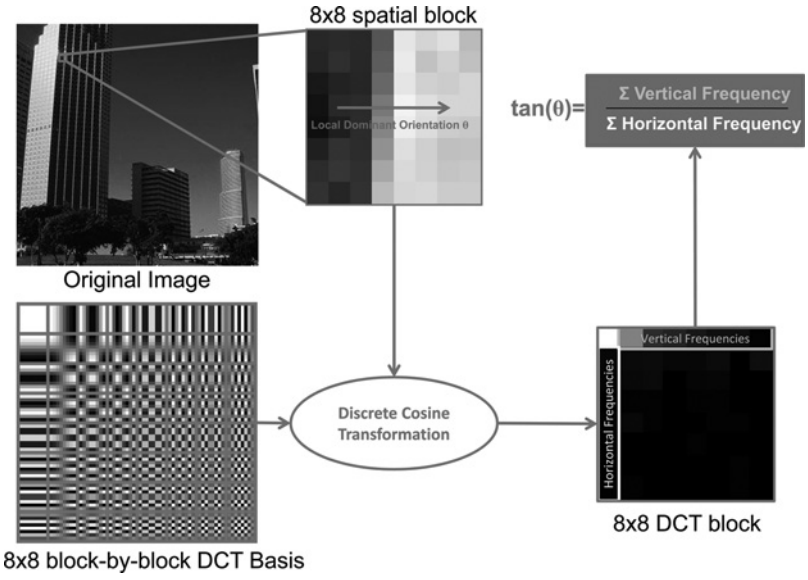

Fig. 2 *LDO extraction process*

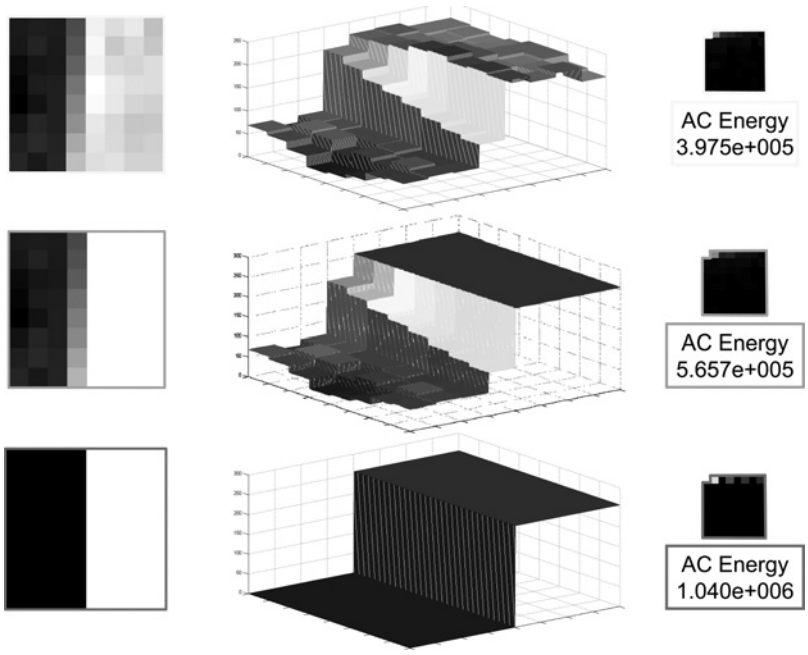

Fig. 3 *Overall AC energy contained on a 8 × 8 DCT block is related to the strength of the LDO of the corresponding 8 × 8 spatial block*

Three 8 × 8 spatial blocks with same LDO and different strength are represented in the left column. The strengths of the blocks (in increasing order from top to bottom) are represented in the middle. The AC energies of each block extracted from DCT domain are reported in the right column. The AC energy increases from top to bottom in accordance with the strengths of the corresponding blocks

The remainder of this paper is organised as follows: in Section 2 related works are reported. Section 3 describes the employed model for image representation and illustrates the data set used to perform experiments. Section 4 details the experimental framework, reporting also the obtained results. Finally, conclusions and avenues for further research are given in Section 5.

2 Related works

Scene recognition is a fundamental process of human vision that allows us to efficiently and rapidly analyse our surroundings. Seminal studies in computational vision [14] have portrayed scene recognition as a progressive reconstruction of the input from local measurements (e.g. edges, 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 concepts and object information [15]. 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 [16]. Recent studies suggested that humans rely on local information as much as on global information to recognise the scene category [17]. In building scene recognition systems some consideration about the spatial envelope properties (e.g. degree of naturalness, degree of openness, etc.) and the level of description (e.g. subordinate, basic, superordinate) of the scene should be taken into account [18].

Different methods have been proposed to build an expressive description of the content of a scene. A wide class of scene recognition algorithms use colour, texture or edge features: Gorkani and Picard [19] used statistics of orientation in the images to discriminate scene into two categories (cities against natural landscapes). Indoor against outdoor classification based on colour, texture and spectral feature (exploiting DCT coefficients) was addressed by Szummer and Picard [20]. Existing methods work extracting local concepts directly on spatial domain [21–25] or in frequency domain [18, 26, 27]. A global representation of the scene may be obtained grouping together these information in different ways. Recently, the spatial layout of the local information [28–30] as well as the metadata information collected during acquisition time have been used to improve the classification quality [31].

In the following, we will elucidate in more details some of the state-of-the-art approaches working with features extracted on frequency domain. A review of state-of-the-art methods working with features extracted on spatial domain can be found in [29, 32].

2.1 Scene classification extracting features on frequency domain

In the last decade different approaches have efficiently exploited the frequency domain as a useful and effective source of information to encode holistically an image for scene classification. The statistics of natural images on frequency domain [33] reveal that there are different spectral signatures for different image categories. In particular by considering the shape of the spectrum of an image, it is possible to address scene category [18, 33, 34], scene depth [35] and object priming [36] such as identity, scale and location.

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 [34] have introduced computational procedures to extract global structural information of complex natural scenes looking at the frequency domain [18, 33, 34]. The computational model that they propose 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 most representative of one of the

doi: 10.1049/iet-cvi.2010.0056

classes and negative for the other. In particular, the sign of the DST values indicates the correlation between the spectral components and the 'spatial envelope' properties of the two classes. When the task is to discriminate between two kinds of scenes (e.g. natural against artificial), a suitable DST is built and used for the classification. A DST is learned in a supervised way using linear discriminant analysis [37]. The classification of a new image can be hence performed by the sign of the correlation between the power spectrum of the 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 are used by Torralba and Oliva for this task. The final classification is performed on the principal component of the sampled frequencies. Oliva and Torralba [18] performed test on a data set containing about 8000 pictures of environmental scenes covering a large variety of outdoor places. Images were 256 × 256 pixels in size, in 256 grey levels. An accuracy of about 94% was obtained on tests performed to discriminate between classes at the superordinate level of description (e.g. natural against artificial).

Luo and Boutell [38] 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 classification. The results obtained with this approach are been similar to results obtained in previous works of Oliva and Torralba.

Farinella *et al.* [26] proposed to exploit features extracted by ordering the discrete Fourier power spectra to capture the naturalness of scenes. By ordering the frequencies spectra, 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 of the discriminative frequencies and the vector of the relative positions in the ordering of the discriminative frequencies are used singularly or in combination to provide a holistic representation of the scene suitable for the classification problem under consideration. An accuracy of 92.6% was obtained just using the ordering position of the selected discriminative frequencies.

The DCT domain was explored by Ladret and Guérin-Dugué [11] to perform image classification and retrieval by using *K*-nearest neighbour (KNN) algorithm. Tests were performed on a database consisting of only 470 pictures (256 × 256 pixels, grey levels values) from COREL database. The images were described at superordinate level of description. Specifically, tests have been done to discriminate natural against artificial scenes (obtaining about 94% of accuracy) as well as to discriminate between four classes of scenes (obtaining about 80.75% of accuracy): outdoor, indoor, closed and open.

The above reported techniques disregard the spatial layout of discriminative frequencies. Torralba *et al.* [35, 36, 39] have proposed to further look at the spatial frequency layout to address more specific vision tasks like object recognition, location detection, scale understanding and scene depth estimation.

Different problems should be considered in transferring the ability of scene recognition to imaging devices domain: limited memory and computational resources. Taking into account the memory constraints in imaging devices, the reviewed approaches cannot be used in their original form. Our work is partially inspired by [11]. We have assessed [11] on a benchmark standard data set used by computer vision community adapting it to work on constrained domain by means of *ad hoc* strategies better specified in the forthcoming sections. We propose to extract features from the DCT blocks just to have useful insight with respect to the overall orientation of the main details present in the image; as described in [18, 26] such info can be used to infer the category of a scene at superordinate level of description. The proposed approach properly works in constrained domain.

3 Histograms of oriented DCT blocks

In this section we introduce the holistic representation used in the proposed classification framework. For sake of simplicity, we focus on the task of natural against artificial classification. The same representation has been used to discriminate between the following classes of scenes: natural against artificial, open against closed, indoor against outdoor, document against landscape and against portraits.

The power spectrum of an image contains enough relevant information about its global structure [11, 18, 26]. As discussed in Section 2.1, different studies pointed out that information extracted in frequency domain can be holistically encoded to represent the scene for naturalness classification. In [18, 33] the authors have experimentally observed, by employing a data set containing different basic classes of scenes category (city, forest, mountain, building, etc.), that the spectrum of natural scenes is quite isotropic with no preferred direction, whereas the spectrum of artificial scenes have strong 'vertical' and 'horizontal' axis. Moreover, experiments in [18, 33] demonstrated that there exist a relation between the spatial envelope of structures into the scenes (such as spatial envelope of edges) and the degree of openness, expansion and roughness. In particular, the spatial envelope of structures into the scenes is useful to discriminate between open against closed classes as well as between indoor against outdoor scenes. In the specific case addressed by this paper, the discrimination between natural against artificial scenes is based on the fact that straight horizontal and vertical lines dominate man-made structures whereas most natural landscapes have textured zones and ondulating contours. Therefore scenes having a distribution of edges (in spatial domain) commonly found in natural landscapes would have a high degree of naturalness, whereas scenes with distribution of edges (in spatial domain) biased towards vertical and horizontal orientations would have a low degree of naturalness. These considerations lead to claim that the distribution of edges' orientations within an image can help to understand the naturalness of the scene under consideration and hence may be used for natural against artificial classification. In this work the term artificial refers to images in which are depicted man-made environments (cities, buildings, streets, etc.), whereas natural refers to images in which natural landscapes are represented (open country, mountain, forest, coast, etc.).

To discriminate between natural and artificial scenes, we build a global representation of the scene after estimating the LDO and strength of each 8 × 8 block belonging to the grey-scale image. Specifically, these information are obtained considering each 8 × 8 block within an image

encoded in the DCT domain. Shen and Sethi [13] proposed the first image-processing approach to extract edge on DCT domain. The method can be used to extract edge information during the JPEG coding of the image (or after) directly in the encoded compressed image domain. Specifically, during JPEG encoding, the image is partitioned into 8×8 pixel-by-pixel non-overlapped blocks transformed by using DCT basis. Each block is then DCT transformed to derive an 8×8 coefficient array, where the (0, 0) element (top-left) is the DC (zero-frequency) component and entries with increasing vertical and horizontal index values represent higher vertical and horizontal spatial frequencies. For each block $B_k(x, y)$ in the original image, the corresponding DCT coefficients $D_k(u, v)$ are generated by using the following equation

$$D_k(u, v) = \alpha(u)\alpha(v) \sum_{x=0}^{7} \sum_{y=0}^{7} B_k(x, y) \cos\left(\frac{\pi(2x+1)u}{16}\right) \times \cos\left(\frac{\pi(2y+1)v}{16}\right) \quad (1)$$

where

$$\alpha(f) = \begin{cases} \frac{1}{\sqrt{8}}, & f = 0 \\ \sqrt{\frac{1}{4}}, & 1 \leq f \leq 7 \end{cases} \quad (2)$$

From this representation, the following equation can be used to obtain the edge orientation of $B_k(x, y)$ [11]

$$\tan(\theta_k) = \frac{\Sigma_{u=1;u=u+2}^{7} D_k(u, 0)}{\Sigma_{v=1;v=v+2}^{7} D_k(0, v)} \quad (3)$$

The local variance of each DCT block (the AC energy) is relative to the strength (Fig. 3) of the edge [13] whose orientation have been evaluated by using (3). The strength can properly weight each edge according to its importance. To evaluate the strength of each block, the following formula is used

$$A_k = \sum_{u=1}^{7} D_k^2(u, 0) + \sum_{v=1}^{7} D_k^2(0, v) + \sum_{u=1}^{7} \sum_{v=1}^{7} D_k^2(u, v) \quad (4)$$

Fig. 4 illustrates the natural (artificial) scene, the LDO of each 8×8 block estimated by using (3) and the relative strength (AC energy) estimated using (4). The arrow lengths on the LDOs plot are proportional to the norm of the vector with component, respectively, equal to the total horizontal energy (5) and the total vertical energy (6) of each 8×8 DCT block.

$$H_k = \sum_{u=1}^{7} D_k^2(u, 0) \quad (5)$$

$$V_k = \sum_{v=1}^{7} D_k^2(0, v) \quad (6)$$

As shown in Figs. 4*a* and *b*, the natural scenes present many horizontal edges (e.g. due to the horizon in the scene), whereas in the artificial scenes the vertical edges are prominent (e.g. due to the buildings in the scene). The strength (AC energy) of each block indicates how much the corresponding LDO should be taken into account. The AC energy (strength) corresponding to homogeneous or texturised image blocks (e.g. sky blocks, sea blocks, clouds blocks, etc.) is lower than the corresponding AC energy (strength) of edge image blocks (e.g. horizon blocks).

A holistic representation of the scene can be built just analysing the distribution of the LDOs weighted taking into account their corresponding strengths [11]. Specifically, for each grey-scale image I coded with K blocks in $\mathrm{DCT}_{8\times8}$ domain, let $\{\theta_1, \ldots, \theta_K\}$ be the K LDOs extracted by using (3) and let $\{A_1, \ldots, A_K\}$ be the K AC energies extracted by using (4). The d-dimensional features vector $\mathrm{LDO}(\mathrm{DCT}_{8\times8}(I)) = [f_{\hat{\theta}_1}, f_{\hat{\theta}_2}, \ldots, f_{\hat{\theta}_d}]^{\mathrm{T}}$ used to represent the whole image I is obtained as follows

$$f_{\hat{\theta}_i} = \frac{N(\hat{\theta}_i)}{SN}, \quad \forall i \in \{1, \ldots, d\} \quad (7)$$

where

- $N(\widehat{\theta_i}) = \sum_{A_k \in \Theta_i} \log(A_k)$
- $\hat{\theta}_i \in [-90, 90]$, $\hat{\theta}_1 = -90$, $\hat{\theta}_{i+1} = \hat{\theta}_i + \frac{180}{d}$, $\hat{\theta}_{d+1} = 90$
- $\Theta_i = \{A_k | \hat{\theta}_i < \theta_k \leq \hat{\theta}_{i+1},\ A_k > \zeta,\ k = 1, \ldots, K\}$
- $\mathrm{SN} = \Sigma_{n=1}^{d} N(\hat{\theta}_n)$ is the normalisation constant
- d is the number of orientation bins
- ζ is a threshold useful to discard the marginal orientations

To effectively build the LDO of an image I, the number of bins to be considered (the parameter d) and the threshold needed to extract only the significant orientations (the parameter ζ) must be fixed. To this aim, the benchmark algorithm KNN and the leave-one-out cross-validation (LOOCV) procedure [37] were employed. Finally, when the best representation parameters have been fixed, we used them in training (offline) the logistic classification model [37] used for the final classification. Finally, the TF-IDF methodology [12] has been investigated to select the most discriminative orientations between different classes of scenes.

Fig. 5 reports the mean LDOs distributions of natural and artificial scenes. The mean LDO distributions in Fig. 5 confirm what was underlined in Section 3: natural scenes present many horizontal edges whereas in the artificial scenes the vertical edges are more evident. The distributions in Fig. 5*a* have been computed averaging the LDO representations of about 1400 natural scenes (Fig. 6): 410 open county, 328 forest, 360 coast and 374 mountain. The distributions in Fig. 5*b* have been computed averaging the LDO representations of about 3200 artificial scenes (Fig. 7): 216 bedrooms, 241 suburban, 311 industrial, 210 kitchens, 289 living rooms, 260 highways, 308 inside city, 292 streets, 356 tall buildings, 215 offices and 315 stores. The aforementioned data set is one of the most complete scene category dataset at basic level of description (15 scene categories [28]). It is an augmented version of the data set used in [18, 22]. All images are considered in grey-scale and encoded in JPEG format with an average size of 244×272 and high variance with respect to the related compression ratio (bit-per-pixel (bpp) $\in$ [0.4169, 11.1396]). The average bpp is 2.3323 ± 1.8621. More information about the composition of the data set are reported in Table 1. This data set has been used to perform

doi: 10.1049/iet-cvi.2010.0056

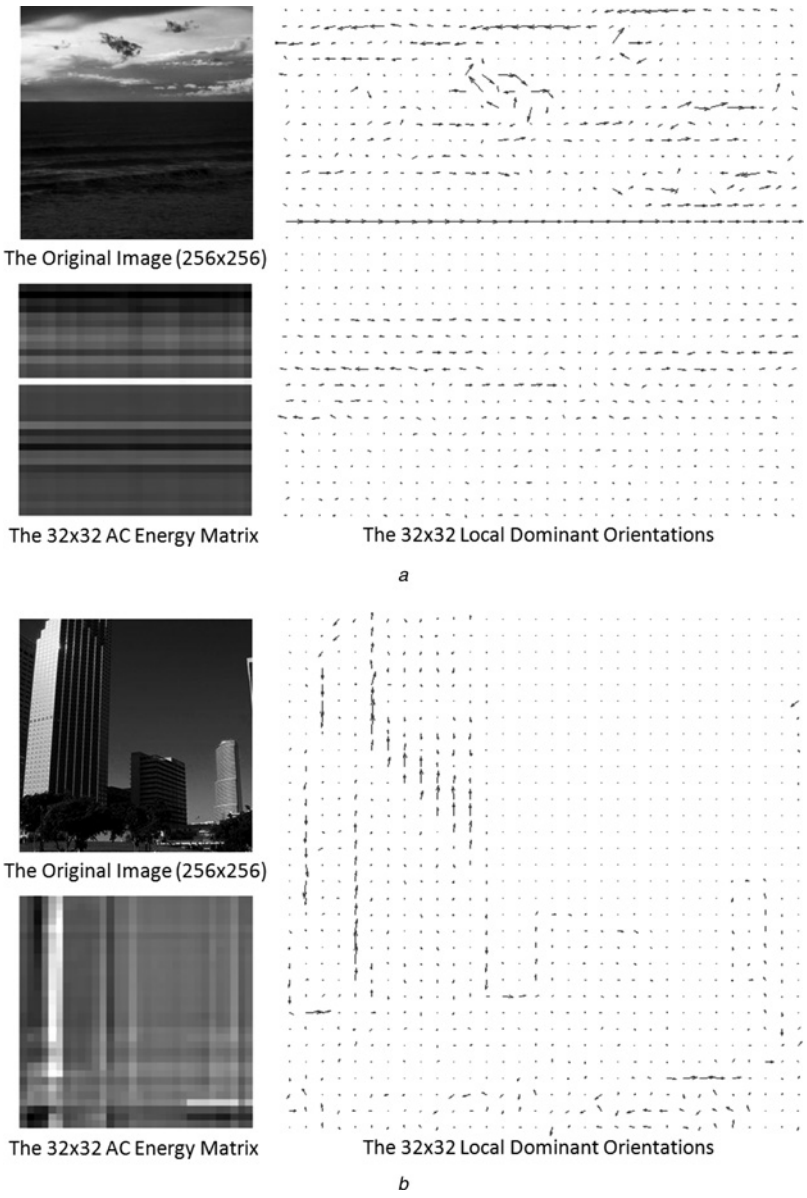

Fig. 4 *LDOs and the corresponding AC energies*

a Natural scene (top-left), the LDO of each 8 × 8 block (right) and the corresponding AC energies (bottom-left)
b Artificial scene (top-left), the LDO of each 8 × 8 block (right) and the corresponding AC energies (bottom-left)

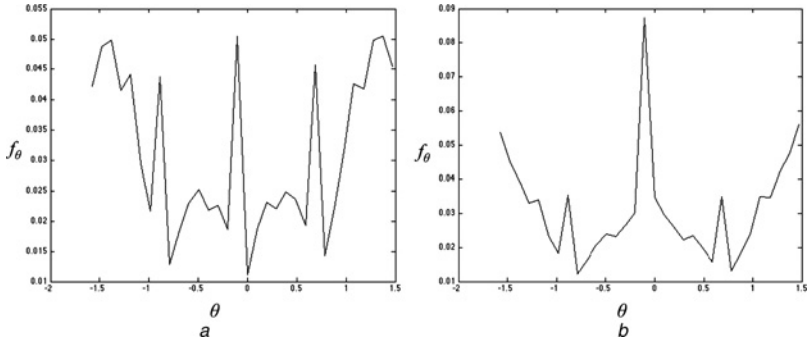

Fig. 5 *Mean LDOs distributions of natural and artificial scenes*

a Mean LDO of natural scenes
b Mean LDO of artificial scenes

experiments reported in the next section. Strongly ambiguous scenes were discarded for testing purposes (e.g. in the case of openness classification, street scenes with no perceived depth have been discarded). The high variability of the compression ratio of the considered data set guarantees that the results obtained with the proposed model are independently from

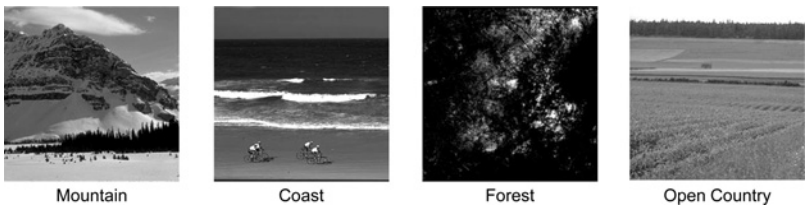

Fig. 6 *Some examples of natural scenes used to compute the average LDOs distributions reported in Fig. 5a*

Fig. 7 *Some examples of artificial scenes used to compute the average LDOs distributions reported in Fig. 5b*

Table 1 Composition of the data set used for evaluation purpose. The data set contains images collected by the authors of [18, 22, 28]

Class	# Images	Average dimension	Average bpp	STD bpp
highway	260	256 × 256	0.9812	0.3015
office	215	220 × 318	1.0410	0.1898
coast	360	256 × 256	1.0751	0.3418
living room	289	224 × 300	1.3318	0.2482
suburban	241	220 × 330	1.5059	0.2286
tall building	356	256 × 256	1.5190	0.3627
mountain	374	256 × 256	1.5667	0.4220
street	292	256 × 256	1.5849	0.2431
open country	410	256 × 256	1.5950	0.4351
inside city	308	256 × 256	1.6709	0.3298
bedroom	216	221 × 289	2.4150	1.9982
forest	328	256 × 256	2.4607	0.5091
kitchen	210	283 × 235	4.4456	1.9408
industrial	311	240 × 286	5.6315	1.0958
store	315	233 × 289	6.5860	0.8729
all classes	4485	244 × 272	2.3323	1.8621

this parameter. In real cases, the application inside a typical IGP (e.g. a constrained domain) can use data of input that are available before quantisation, also for multiple classes as reported in Section 4.6. Moreover, if the proposed method is used after acquisition (and coding), it is important to highlight that in high-end cameras the quantisation factor is usually very low.

The classification engine we propose is based on the differences concerning the 'shape' of the LDO distributions (Fig. 5), which is also effective to discriminate between different classes of scenes as pointed out in the following section.

4 Experiments and results

First of all we present a series of experiments aimed to justify the parameters setting of the proposed holistic representation of the scene (e.g. number of bins histogram of oriented blocks, exploitation of TF-IDF methodology, etc.). Such settings are then used to represent the images together with a simple and efficient probabilistic classifier. Specifically, we use a KNN [37] to fix the best representation parameters, whereas a logistic classification model [37] is trained and used as final framework for classification.

All the experiments that involve KNN for class discrimination make use of the LOOCV procedure [37]. The final classification results are obtained averaging on the results of all the LOOCV runs. The following parameters have been involved in the experiments: the number d of orientation bins, the strength threshold ζ, the similarity measure S used by KNN and the number of neighbours K used in the nearest neighbour rule. Each experiment was related to the evaluation of the classification performance by using a point into the parameter space $d \times \zeta \times S \times K$ (properly sampled in a grid of 960 points). The experiments pointed out that the best parameters to be used when the proposed holistic representation is coupled with KNN are

doi: 10.1049/iet-cvi.2010.0056

the following: $d = 32$, ζ equal to 10% of the maximal A_k extracted from the image I under consideration, the similarity measure based on Bhattacharyya coefficient, and $K = 3$. In the next subsections, we report a subset of the experiments in which after having three fixed parameters the remaining parameter has been evaluated with respect to the classification rate.

All experiments in which a logistic classifier is involved to discriminate between classes of scenes are repeated ten times with different randomly selected training (50%) and test images (50%). The parameters involved in the classification models have been learned at each run from the training set through classic maximum likelihood estimation (MLE) procedure and the confusion matrices were recorded at each run. The classification results are obtained averaging on the results of all ten runs.

4.1 Number of orientation bins

Fig. 8 reports the natural against artificial classification accuracy results with respect to the parameter d involved in (7). This parameter specifies the number of orientation bins to be considered in building the holistic representation of the scene $\mathrm{LDO}(\mathrm{DCT}_{8\times8}(I)) = [f_{\hat{\theta}_1}, f_{\hat{\theta}_2}, \ldots, f_{\hat{\theta}_d}]^{\mathrm{T}}$. All remaining parameters have been fixed as reported at the beginning of Section 4. The TF-IDF methodology [12] was used in representing the scenes. The metric based on Bhattacharyya coefficient [40, 41] have been employed in the KNN classifier (e.g. $K = 3$). As shown in Fig. 8, by using a very compact representation of the scene composed of $d = 32$ orientation bins, the classification results are higher than 94%. Table 2 reports the average confusion matrix obtained averaging on the overall leave-one-out runs in which a representation with $d = 32$ orientation bins have been used (the x-axis represents the inferred classes while the y-axis represents the ground-truth category).

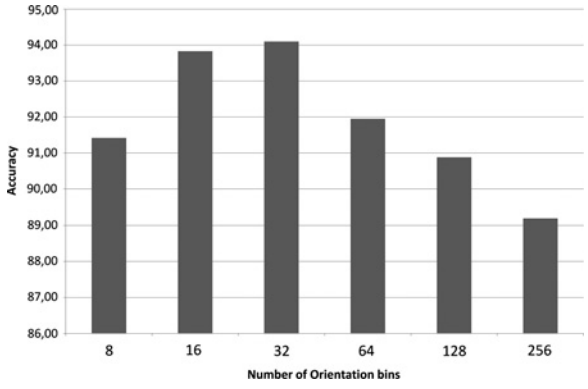

Fig. 8 *Natural against artificial classification accuracy with respect to the number of orientation bins*

Best results have been obtained by using a compact representation of 32 bins

Table 2 Confusion matrix considering a representation with $d = 32$ orientations bins

	Natural	Artificial
Natural	94.15	5.85
Artificial	5.94	94.06

Average classification rates for the natural and artificial classes are listed along the diagonal. The average accuracy is 94.10%

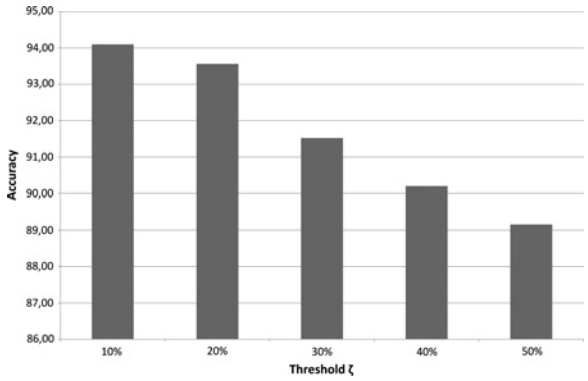

Fig. 9 *Natural against artificial classification accuracy with respect to the threshold used to select the significant orientations*

Best results are obtained when the threshold ζ is set to be equal to 10% of the maximal A_k extracted from the image I under consideration

4.2 Threshold on strength to select significant orientations

These experiments were performed to establish the best strength threshold ζ to be used in building the representation of the scene. The number of orientation bins was fixed to $d = 32$. All the other parameters have been fixed as reported at the beginning of Section 4. In Fig. 9 the natural against artificial classification results with respect to the different threshold are reported. The best results are obtained when ζ is equal to 10% of the maximal A_k extracted from the image I under consideration during the representation phase. Increasing this threshold, significant orientations are discarded and the classification accuracy decreases. Values of ζ less than 10% of the maximal A_k extracted from the image I under consideration have not been able to capture the concept of 'edge'.

4.3 Similarity between local orientation distributions

The similarity measure used in the KNN classifier has a clear impact in the classification results. We tested different similarity measures taking into account that the classification should be based on the LDO 'shape' similarities/differences. Let $L^{(1)} = \mathrm{LDO}(\mathrm{DCT}_{8\times8}(I^{(1)})) = [f^{(1)}_{\hat{\theta}_1}, f^{(1)}_{\hat{\theta}_2}, \ldots, f^{(1)}_{\hat{\theta}_d}]^{\mathrm{T}}$ and $L^{(2)} = \mathrm{LDO}(\mathrm{DCT}_{8\times8}(I^{(2)})) = [f^{(2)}_{\hat{\theta}_1}, f^{(2)}_{\hat{\theta}_2}, \ldots, f^{(2)}_{\hat{\theta}_d}]^{\mathrm{T}}$, the LDOs distributions of two different images $I^{(1)}$ and $I^{(2)}$. We tested the following similarity measures: absolute difference (8), metric based on Bhattacharyya coefficient (9), χ^2 distance (10), Jeffrey divergence (11), Pearson correlation coefficient (12), square difference distance (13) and weighted Euclidean distance (14). The distance based on Bhattacharyya coefficient has several desirable properties [41]: (i) it imposes a metric structure, (ii) has a geometric interpretation since it is related to the cosine of the angle between two vectors, (iii) is valid for arbitrary distributions and (iv) approximates the χ^2 distance, while avoiding the singularity problem of the χ^2 test when comparing empty histogram bins.

$$S(L^{(1)}, L^{(2)}) = \sum_{n=1}^{d} |f^{(1)}_{\hat{\theta}_n} - f^{(2)}_{\hat{\theta}_n}| \tag{8}$$

$$S(L^{(1)}, L^{(2)}) = \sqrt{1 - \sum_{n=1}^{d} \sqrt{f_{\hat{\theta}_n}^{(1)} * f_{\hat{\theta}_n}^{(2)}}} \quad (9)$$

$$S(L^{(1)}, L^{(2)}) = \sum_{n=1}^{d} \frac{(f_{\hat{\theta}_n}^{(1)} - m_i)^2}{m_i} \quad (10)$$

where $m_i = (f_{\hat{\theta}_n}^{(1)} + f_{\hat{\theta}_n}^{(2)}/2)$.

$$S(L^{(1)}, L^{(2)}) = \sum_{n=1}^{d} \left\{ \left[f_{\hat{\theta}_n}^{1} * \log\left(\frac{f_{\hat{\theta}_n}^{(1)}}{m_i} \right) \right] + \left[f_{\hat{\theta}_n}^{(2)} * \log\left(\frac{f_{\hat{\theta}_n}^{(2)}}{m_i} \right) \right] \right\} \quad (11)$$

where $m_i = (f_{\hat{\theta}_n}^{(1)} + f_{\hat{\theta}_n}^{(2)}/2)$.

$$S(L^{(1)}, L^{(2)}) = 1 - \frac{\sum_{n=1}^{d} ((f_{\hat{\theta}_n}^{(1)} - \mu^{(1)}/\sigma^{(1)}) * (f_{\hat{\theta}_n}^{(2)} - \mu^{(2)}/\sigma^{(2)}))}{d} \quad (12)$$

where $\mu^{(1)}$ and $\sigma^{(1)}$ are, respectively, mean and standard deviation of the vector $L^{(1)}$, whereas $\mu^{(2)}$ and $\sigma^{(2)}$ are, respectively, mean and standard deviation of the vector $L^{(2)}$.

$$S(L^{(1)}, L^{(2)}) = \sum_{n=1}^{d} (f_{\hat{\theta}_n}^{(1)} - f_{\hat{\theta}_n}^{(2)})^2 \quad (13)$$

$$S(L^{(1)}, L^{(2)}) = \sum_{n=1}^{d} w_n * (f_{\hat{\theta}_n}^{(1)} - f_{\hat{\theta}_n}^{(2)})^2 \quad (14)$$

where $w_n = f_{\hat{\theta}_n}^{(1)}$ if $f_{\hat{\theta}_n}^{(1)} \neq 0$, $w_n = 1$ otherwise.

In evaluating the performances with respect to the similarity measure involved in the KNN (e.g. $K = 3$), we fixed the number of orientation bins and the strength threshold as reported at the beginning of Section 4. The TF-IDF methodology was used in representing the scene. As shown in Table 3, the best results are obtained when the metric based on Bhattacharyya coefficient is employed. We finally performed experiments in order to establish the number of neighbours to use in the KNN rule. Fig. 10 shows the results obtained using KNN coupled with the metric based on Bhattacharyya coefficient and different number of neighbours ($K = 1, 3, 5, 7$). The best results have been obtained with $K = 3$.

Table 3 Natural against artificial classification accuracy with respect to the similarity measure used in the KNN algorithm

Similarity measure	Accuracy
Absolute difference	91.68
Bhattacharyya	94.10
χ^2	93.83
Jeffrey	89.54
Pearson	68.63
Square difference	90.08
Weighted Euclidean	89.54
Fourier distance	93.02

Metric based on Bhattacharyya coefficient outperforms the other similarity measures in terms of classification accuracy

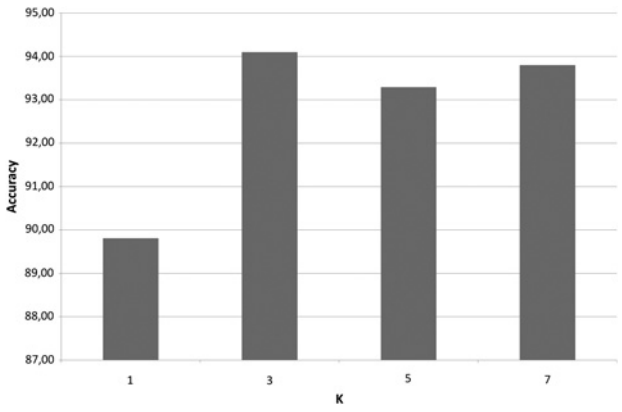

Fig. 10 *Classification results obtained using KNN with the metric based on Bhattacharyya coefficient and different number of neighbours (K = 1, 3, 5, 7)*

Note that, among the similarity measures used in these experiments, we also tested the Euclidean distance on the real Fourier coefficient ('Fourier distance') as previously suggested by Ladret and Guérin-Dugué [11]. The local dominant distribution is treated as a discrete signal and the Fourier transformation is employed. The Euclidean distance between the first two components is used to establish the similarity between scenes. The experiments pointed out that the similarity measure based on Bhattacharyya coefficient proposed in this paper outperforms the similarity measures proposed in [11] (see Table 3).

4.4 LDO against TF-IDF-LDO

All of the experiments above have been performed by using the TF-IDF methodology [12]. In this case the TF-IDF methodology is applied to select the most discriminative orientations between natural and artificial classes. Tables 4 and 5 show the confusion matrices obtained when the representation of the scene is obtained with or without the TF-IDF methodology. Despite good results being obtained using local dominant distribution alone (Table 4: 91.78%), the experiments demonstrate that the exploitation of TF-IDF methodology improves the classification accuracy (Table 5: 94.10%).

Table 4 Confusion matrix considering TF-IDF-LDO representation

	Natural	Artificial
Natural	94.15	5.85
Artificial	5.94	94.06

Average classification rates for the natural and artificial classes are listed along the diagonal. The average accuracy is 94.10%

Table 5 Confusion matrix considering LDO representation

	Natural	Artificial
Natural	91.1	7.9
Artificial	7.53	92.47

Average classification rates for the natural and artificial classes are listed along the diagonal. The average accuracy is 91.78%

doi: 10.1049/iet-cvi.2010.0056

Fig. 11 reports some examples of images classified employing a KNN and the similarity measure based on Bhattacharyya coefficient (9). In particular, the images to be classified are depicted in the first column, whereas the first three closest images used to establish the proper class of test image are reported in the remaining columns. The results are semantically consistent in terms of visual content (and category) to the related images to be classified.

4.5 KNN against logistic classification

As pointed out by the experiments in previous sections, a compact 32-dimensional vector can be used to holistically represent the scene. Although effective results are obtained when KNN is used as classifier, its implementation in domain with limited time, space and computational power resources (e.g. digital camera, mobile phone, etc.) is not straightforward. Indeed, KNN is a memory-based classifier (e.g. the training set must be taken in memory for classification purpose) whose computational costs and space resources exceed the constrained domain of our interest. To overcome these difficulties, we propose to use a logistic model for classification purpose [37]. The basic assumption is that the difference between the logarithms of the class-conditional density functions is linear in the vector $\boldsymbol{f}$ representing the images through TF-IDF-LDO

$$\begin{aligned}&\log\left(P(\boldsymbol{f}|\text{Artificial})\right)-\log\left(P(\boldsymbol{f}|\text{Natural})\right)\\&\quad=w_0+w_1f_{\hat{\theta}_1}+\cdots+w_df_{\hat{\theta}_d}\end{aligned}\tag{15}$$

Such basic assumption is equivalent to [37]

$$P(\text{Natural}|\boldsymbol{f})=\frac{1}{1+e^{(w_0'+w_1f_{\hat{\theta}_1}+\cdots+w_df_{\hat{\theta}_d})}}\tag{16}$$

$$P(\text{Artificial}|\boldsymbol{f})=\frac{e^{(w_0'+w_1f_{\hat{\theta}_1}+\cdots+w_df_{\hat{\theta}_d})}}{1+e^{(w_0'+w_1f_{\hat{\theta}_1}+\cdots+w_df_{\hat{\theta}_d})}}\tag{17}$$

where

$$w_0'=w_0+\log\left(\frac{P(\text{Artificial})}{P(\text{Natural})}\right)$$

In our experiments we assume equiprobability for the priors P(Natural) and P(Artificial), hence $w_0'=w_0$. It is interesting to note that the decision about discrimination between natural against artificial scenes is determined solely by the following linear function

$$g(f)=w_0'+w_1f_{\hat{\theta}_1}+\cdots+w_df_{\hat{\theta}_d}\tag{18}$$

Specifically, a new observation is assigned to the class natural if the value of the function (18) is negative, otherwise is assigned to the class artificial. Hence, after learning the parameters of the logistic classification model (out of the devices), a simple evaluation of a linear function can be used for classification purpose; this leads to overcome the difficulties due to constrained domain of consumer imaging devices.

The experiments performed through KNN pointed out that the best representation of the scene is encoded in a 32-dimensional vector. In our experiments we learn a 33-dimensional vector relative to the parameters involved in the logistic classifier. The parameters vectors may be estimated (e.g. out of the devices) using a classic maximum likelihood estimation approach [37].

After that the learning phase used to train the model is concluded, a new image can be assigned to a class $\hat{c}\in\{\text{Natural},\ \text{Artificial}\}$ according to a MAP rule

$$\hat{c}=\arg\max_c P(c|f,\widehat{w}_0',\ \ldots,\widehat{w}_d)\tag{19}$$

In Table 6 the classification results obtained using the logistic classification model are reported. The average accuracy is 93.01%. Although the recognition results are about 1.09% less than the results obtained by using KNN (see Table 4 for comparison of KNN against logistic results), one should not overlook that the proposed method outperforms KNN in constrained domain (e.g. limited resources in terms of space, time and computational power). Note that the obtained results match in accuracy with the work presented in [11] by reducing the computational resources needed for the classification task.

Fig. 11 *Examples of images classified by using the proposed image representation, KNN and the similarity measure based on Bhattacharyya coefficient*

Target images are on the top row and the three closest images are shown on the bottom rows. Each column corresponds to a different class of scene

Table 6 Confusion matrix obtained through logistic classification

	Natural	Artificial
Natural	93.03	6.97
Artificial	6.70	93.30

Average classification rates for the natural and artificial classes are listed along the diagonal. The average accuracy is 93.01%

4.6 Classification performances on other classes of scenes

The proposed classification framework can be applied to other classes of scene. In this section we show preliminary results obtained by applying the proposed approach on different opposite classes of scene at superordinate level of description: open against closed, indoor against outdoor. These classes may be useful to properly address some parameters of IGP employed within imaging devices [1, 6]. Moreover, we present a simple extension of the proposed method to work with multiple classes by just considering three classes usually managed in some way by a digital camera or a mobile phone: landscape, document and portraits. All the experiments of this section have been done employing the logistic classification model with the same parameters pointed out in previous sections.

First, let us examine the performance of the proposed approach to discriminate open against closed scenes. A closed scene is a scene with small perceived depth, whereas an open scene is a scene with big perceived depth. The database used for open against closed classification experiments is composed of eight basic scene categories collected by [18]: coast (360 images), open county (410 images), street (292 images), highway (260 images), forest (328 images), mountain (374 images), tall building (356 images) and city (308 images). The first four basic classes have been considered as open scenes (1322 images), whereas the other classes have been considered as closed scenes (1366 images). The images were considered in grey-scale since colour information can be discarded in building the proposed image representation. Some examples of open and closed scenes are given in Fig. 12. The experiments on this data set are repeated 20 times with different randomly selected training (75%) and test (25%) images. The parameters involved in the logistic classification model have been learned at each run from the training set through classic MLE procedure (see Section 4.4) and the confusion matrices were recorded at each run. The final open against closed classification results are obtained averaging on the results of all 20 runs.

Fig. 13 reports the polar version of the mean LDOs distributions of open and closed scenes. The distributions in Fig. 13*a* have been computed averaging the LDO

Fig. 12 *Some examples of open and closed scenes used to compute the average LDOs distributions reported in Fig. 13*

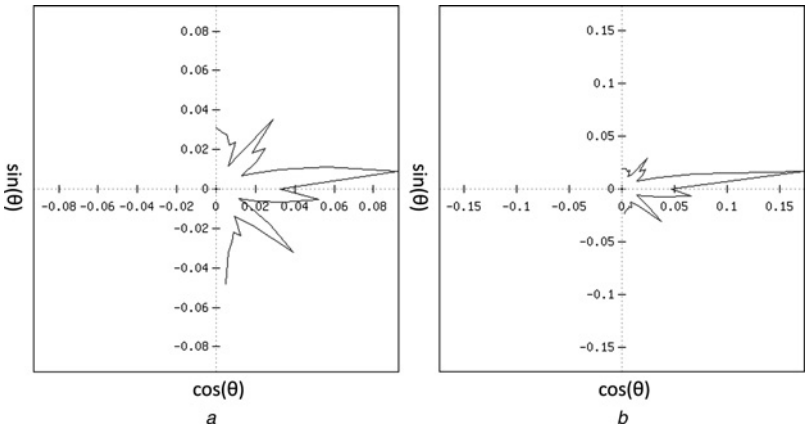

Fig. 13 *Polar version of the mean LDOs distributions of open and closed scenes*

a Mean polar LDO of open scenes
b Mean polar LDO of closed scenes

doi: 10.1049/iet-cvi.2010.0056

representations of the open scenes of the data set described above. The distributions in Fig. 13*b* have been computed averaging the LDO representations of the closed scenes of the data set described above. As shown in Fig. 13, different orientations are represented in the mean local orientations distribution of open class, whereas the vertical edges are more frequent in the mean local orientations distribution of closed class. In Table 7 the classification results obtained on open against closed classification are reported. The experiments pointed out that the proposed approach can be effectively used to discriminate between opposite classes at superordinate level of description differently than natural against artificial.

Next, let us examine the performances of in against out classification. The database used for these experiments is composed of nine basic scene categories collected by [18, 22]: coast (360 images), highway (260 images), mountain (374 images), open county (410 images), street (292 images), bedroom (216 images), kitchen (210 images), living room (289 images) and office (215 images). The first five basic classes have been considered as outdoor scenes (1669 images), whereas the other classes have been considered as indoor scenes (930 images). Some examples of indoor and outdoor scenes used in our experiments are given in Fig. 14. Also in this case, the images were considered in grey-scale. The experiments on this data set are repeated 20 times with different randomly selected training (75%) and test (25%) images. The parameters involved in the logistic classification model have been learned at each run from the training set through classic MLE procedure (see Section 4.4) and the confusion matrices were recorded at each run. The final indoor against outdoor classification results are obtained averaging on the results of all 20 runs.

Table 7 Open against closed classification results

	Open	Closed
Open	89.56	10.44
Closed	8.65	91.35

Average classification rates for the open and closed classes are listed along the diagonal. The average accuracy is 90.33%

Fig. 15 reports the polar version of the mean LDOs distributions of outdoor and indoor scenes. The distributions in Fig. 15*a* have been computed averaging the LDO representations of the outdoor scenes of the data set described above. The distributions in Fig. 15*b* have been computed averaging the LDO representations of the indoor scenes of the data set described above. Also, in this case the 'shapes' of the mean LDOs of the two involved classes are quite different: vertical edges are more frequent in the mean local orientations distribution of indoor class. In Table 8 the classification results obtained on indoor against outdoor classification are reported (90.89% accuracy).

The proposed method for indoor against outdoor classification outperforms the approach proposed by Szummer *et al.* [20] both in terms of computational resources and classification accuracy. In [20] the best indoor

Fig. 14 *Some examples of outdoor and indoor scenes used to compute the average LDOs distributions reported in Fig. 15*

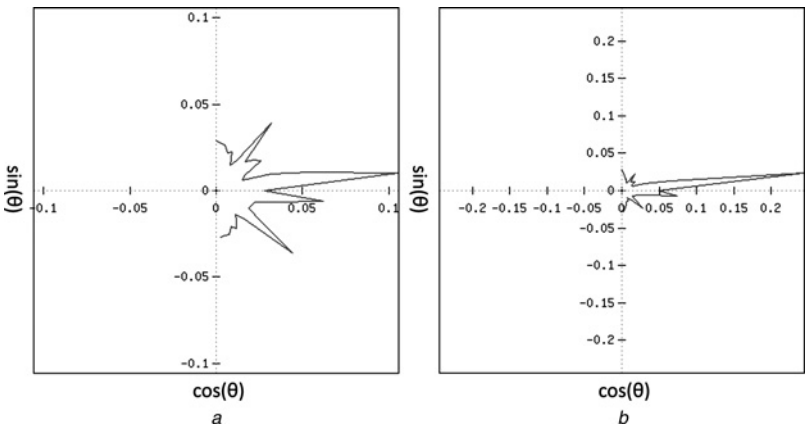

Fig. 15 *Polar version of the mean LDOs distributions of outdoor and indoor scenes*

a Mean polar LDO of outdoor scenes
b Mean polar LDO of indoor scenes

Table 8 Outdoor against indoor classification results

	Outdoor	Indoor
Outdoor	90.39	9.61
Indoor	8.60	91.40

Average classification rates for the outdoor and indoor classes are listed along the diagonal. The average accuracy is 90.89%

against outdoor classification results (90.3% accuracy) have been obtained by using KNN with $K = 13$ and multiple features (colour and MSAR features). Despite DCT features have been tested to encode textures by Szummer *et al.* [20], the best results obtained exploiting DCT features stated at 84% using a two-stage classification algorithm involving KNN with $K = 13$.

Finally, let us examine a simple extension of the proposed approach to work with multiple classes. Specifically, we considered the problem of discriminating between three classes usually acquired by a digital camera: document, landscape and portrait. Scenes belonging to these three classes are usually acquired by a consumer imaging device; the class inferred through a scene recognition engine may be useful for different tasks within IGP (e.g. colour constancy [6], optimised compression [9], etc.).

The database used in our experiments is composed of 1532 colour images with different resolution and compression factor. The data set is not freely available due to an NDA with STMicroelectronics. In particular, 382 images are landscape scenes, 874 images are document scenes and 276 images are portraits scenes. Some examples of images that have been used in our experiments are depicted in Fig. 16.

All the images were preprocessed to be considered in 256×256 grey-scale thumbnail version. The 32×32 DCT blocks of each thumbnail have been used to build the TF-IDF-LDO representation as described in Section 3. Also, in this case we used $d = 32$ orientation for the LDO representation, and a strength threshold ζ equal to 10% of the maximal A_k extracted from the image I under consideration.

Fig. 17 reports the polar version of the mean LDOs distributions of the three involved classes of scenes. The distributions in Fig. 17 have been computed averaging the LDO representations of the document, landscape and portrait scenes of the data set. In this case, the mean LDO of document class looks similar to the distribution of edges within an artificial scene ('vertical' edges are more represented), the mean LDO of landscape class looks similar to the edges distribution of natural scenes, whereas the mean LDO of portrait scene differs from the LDOs of the other two classes. To perform our experiments on the $N = 3$ classes of scenes mentioned above, we employed the one-against-all method [37]. This method constructs N binary classifiers (e.g. N binary logistic classifiers). The nth classifier is trained to discriminate samples in class C_n (the positive class) from those in the remaining classes (the negative class). Thus, using a logistic model as binary

Fig. 16 *Some examples of document, landscape and portrait scenes used to compute the average LDOs distributions reported in Fig. 17*

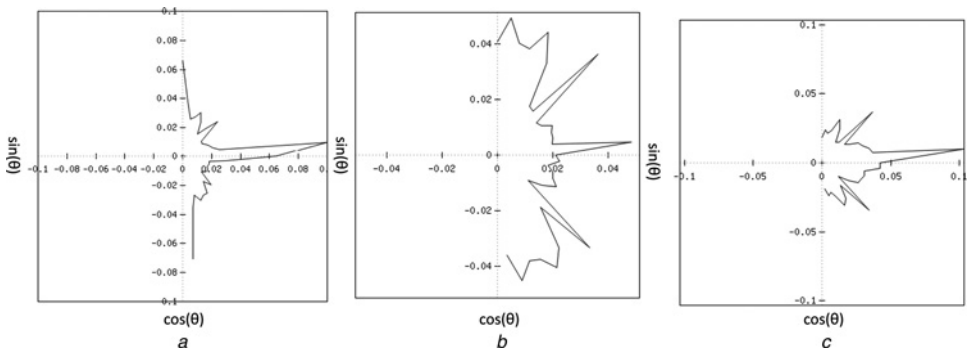

Fig. 17 *Polar version of the mean LDOs distributions of document, landscape and portrait scenes*

a Mean LDO of document scenes
b Mean LDO of landscape scenes
c Mean LDO of portrait scenes

doi: 10.1049/iet-cvi.2010.0056

Table 9 Classification results obtained applying the proposed approach on the three classes usually acquired by an imaging device: portrait, landscape and document

	Portrait	Landscape	Document
Portrait	94.15	2.05	3.8
Landscape	4.34	87.00	8.66
Document	10.18	8.09	81.73

Average classification rates for the three classes are listed along the diagonal. The average accuracy is 87.62%

classifier, after the training phase of all N binary classifiers through classic MLE procedure, the corresponding N binary discrimination functions (20) are evaluated to establish the class of a new sample $\boldsymbol{f}$.

$$g_n(\boldsymbol{f}) = w'_{n,0} + w_{n,1} f_{\hat{\theta}_1} + \cdots + w_{n,d} f_{\hat{\theta}_d}, \quad n = 1, \ldots, N \tag{20}$$

Ideally, for a given sample $\boldsymbol{f}$, the quantity $g_n(\boldsymbol{f})$ will be positive for one value of n and negative for the reminder, giving a clear indication of the class. If there is more than one class for which the quantity $g_n(\boldsymbol{f})$ is positive, the $\boldsymbol{f}$ sample may be assigned to the class $\widehat{c} \in \{C_1, C_2, \ldots, C_N\}$ for which the distance to the hyperplane is the largest (21).

$$\widehat{c} = \arg\max_n \left(\frac{g_n(\boldsymbol{f})}{||\boldsymbol{w}_n||} \right) \tag{21}$$

If all the values of $g_n(\boldsymbol{f})$ are negative, then the $\boldsymbol{f}$ sample is assigned to the class with smallest distance to the hyperplane (22).

$$\widehat{c} = \arg\min_n \left(\frac{g_n(\boldsymbol{f})}{||\boldsymbol{w}_n||} \right) \tag{22}$$

The experiments have been repeated 20 times with different randomly selected training (50%) and test (50%) images. The parameters involved in the multi-class logistic model have been learned at each run from the training set and the confusion matrices were recorded at each run. The final classification results are obtained averaging on the results of all 20 runs. Table 9 reports the confusion matrix obtained averaging on the results of all 20 runs. The average accuracy is 87.62%. Note that the proposed approach is able to work on thumbnail version of the acquired images; this is useful to save computational time in constrained domain.

5 Summary, conclusions and future works

The problem of scene classification is currently of great interest for research community. Moreover, a software engine able to automatically infer information about the category of a scene is useful to support IGP in domains that have limited resources in terms of space, time and computational power, such as consumer imaging devices (e.g. digital still camera, mobile imaging devices, etc.). Infer the class of a scene can be effectively used to optimise IGP domain in both pre-capture phase (e.g. autofocus, auto exposure, white balance) and post-acquisition processing (e.g. colour rendering or scene-based optimisation coding).

The main aim of our work has been to build a scene recognition engine to discriminate between different classes of scenes taking into account the involved constraints. The proposed approach can be directly implemented in the DCT domain, fully compatible with the JPEG format. The holistic representation of the scene is coded with simple operations in a very compact low-dimensional vector. A simple probabilistic model is used for classification purpose. The model's parameters can be learned offline (out of the device) or can be used online to classify a new observed scene through a simple decision function. A very compact low-dimensional vector of parameters is maintained to perform classification. Despite the proposed approach works on constrained domain, the performances of natural against artificial classification closely match the other state-of-the-art methods working on frequency domain. Moreover, as demonstrated by experiments, the proposed method can be extended to work with classes different from natural against artificial and to multiple classes. Further investigation should be devoted in exploiting LDO representation with other kinds of features (e.g. colour, texture, etc.) and EXIF information.

6 Acknowledgments

The work described in this paper was supported by Advanced System Technology Group of STMicroelectronics.

7 References

1 Lukac, R.: 'Single-sensor imaging: methods and applications for digital cameras' (CRC Press, 2008)
2 Battiato, S., Bruna, A.R., Messina, G., Puglisi, G.: 'Image processing for embedded devices' (Bentham Science Publisher, 2010)
3 Zhang, L., Wu, X., Zhang, D.: 'Color reproduction from noisy CFA data of single sensor digital cameras', *IEEE Trans. Image Process.*, 2007, **16**, (9), pp. 2184–2197
4 Zhang, L., Lukac, R., Wu, X., Zhang, D.: 'PCA-based spatially adaptive denoising of CFA images for single-sensor digital cameras', *IEEE Trans. Image Process.*, 2009, **18**, (4), pp. 797–812
5 Battiato, S., Bosco, A., Bruna, A.R., Rizzo, R.: 'Noise reduction for CFA image sensors exploiting HVS behavior', *Sens. J. – MDPI Open Access – Special Issue on Integrated High-Performance Imagers*, 2009, **9**, (3), pp. 1692–1713
6 Bianco, S., Ciocca, G., Cusano, C., Schettini, R.: 'Improving color constancy using indoor–outdoor image classification', *IEEE Trans. Image Process.*, 2008, **17**, (12), pp. 2381–2392
7 Battiato, S., Castorina, A., Mancuso, M.: 'High dynamic range imaging for digital still camera: an overview', *SPIE J. Electron. Imaging*, 2003, **12**, (3), pp. 459–469
8 Battiato, S., Bosco, A., Castorina, A., Messina, G.: 'Automatic image enhancement by content dependent exposure correction', *EURASIP J. Appl. Signal Process.*, 2004, **12**, (3), pp. 1849–1860
9 Battiato, S., Mancuso, M., Bosco, A., Guarnera, M.: 'Psychovisual and statistical optimization of quantization tables for DCT compression engines'. IEEE Int. Conf. on Image Analysis and Processing, 2001, pp. 602–606
10 Nikon Corporation: 'Scene recognition system for more accurate autofocus, auto exposure and auto white balance'. Available at: http://www.nikon.com/, 2007
11 Ladret, P., Guérin-Dugué, A.: 'Categorisation and retrieval of scene photographs from jpeg compressed database', *Patt. Anal. Appl.*, 2001, **4**, (2–3), pp. 185–199
12 Salton, G., Buckley, C.: 'Term-weighting approaches in automatic text retrieval', *Inf. Process. Manage.*, 1988, **24**, (5), pp. 513–523
13 Shen, B., Sethi, I.K.: 'Direct feature extraction from compressed images', *Storage Retrieval Image Video Databases (SPIE)*, 1996, **2670**, pp. 404–414
14 Marr, D.: 'Vision: a computational investigation into the human representation and processing of visual information' (W.H. Freeman and Company, 1982)
15 Biederman, I.: 'Aspects and extension of a theory of human image understanding', in Pylyshyn, Z. (Ed.): 'Computational processes in

human vision: an interdisciplinary perspective' (Ablex Publishing Corp., 1988)
16 Oliva, A., Torralba, A.: 'Building the gist of a scene: the role of global image features in recognition', *Progr. Brain Res.*, 2006, **155**, (1), pp. 251–256
17 Vogel, J., Schwaninger, A., Wallraven, C., Bülthoff, H.H.: 'Categorization of natural scenes: local versus global information and the role of color', *ACM Trans. Appl. Percept.*, 2007, **4**, (3), pp. 1–21
18 Oliva, A., Torralba, A.: 'Modeling the shape of the scene: a holistic representation of the spatial envelope', *Int. J. Comput. Vis.*, 2001, **42**, (3), pp. 145–175
19 Gorkani, M., Picard, R.: 'Texture orientation for sorting photos at a glance'. Int. Conf. on Pattern Recognition, 1994, pp. 459–464
20 Szummer, M., Picard, R.W.: 'Indoor–outdoor image classification'. IEEE Int. Workshop on Content-based Access of Image and Video Databases, 1998, pp. 42–51
21 Vogel, J., Schiele, B.: 'Semantic modeling of natural scenes for content-based image retrieval', *Int. J. Comput. Vis.*, 2007, **72**, (2), pp. 133–157
22 Fei-Fei, L., Perona, P.: 'A Bayesian hierarchical model for learning natural scene categories'. IEEE Int. Conf. on Computer Vision and Pattern Recognition, 2005, pp. 524–531
23 Renninger, L.W., Malik, J.: 'When is scene recognition just texture recognition?', *Vis. Res.*, 2004, **44**, (19), pp. 2301–2311
24 Schettini, R., Brambilla, C., Ciocca, G., Valsasna, A., de Ponti, M.: 'A hierarchical classification strategy for digital documents', *Patt. Recogn.*, 2002, **35**, (35), pp. 1759–1769
25 Turtinen, M., Pietikäinen, M.: 'Visual training and classification of textured scene images'. Int. Workshop on Texture Analysis and Synthesis, 2003, pp. 101–106
26 Farinella, G.M., Battiato, S., Gallo, G., Cipolla, R.: 'Natural versus artificial scene classification by ordering discrete Fourier power spectra'. Joint IAPR Int. Workshop on Structural, Syntactic, and Statistical Pattern Recognition, 2008, pp. 137–146
27 Battiato, S., Farinella, G.M., Gallo, G., Messina, E.: 'Naturalness classification of images into DCT domain'. SPIE – IS&T 21th Annual Symp. on Electronic Imaging Science and Technology – Digital Photography V, 2009, pp. 1–12
28 Lazebnik, S., Schmid, C., Ponce, J.: 'Beyond bags of features: spatial pyramid matching for recognizing natural scene categories'. IEEE Int. Conf. on Computer Vision and Pattern Recognition, 2006, pp. 2169–2178
29 Bosch, A., Zisserman, A., Muñoz, X.: 'Scene classification using a hybrid generative/discriminative approach', *IEEE Trans. Patt. Anal. Mach. Intell.*, 2008, **30**, (4), pp. 712–727
30 Battiato, S., Farinella, G.M., Gallo, G., Ravì, D.: 'Exploiting textons distributions on spatial hierarchy for scene classification', *EURASIP J. Image Video Process.*, 2010, **2010**, pp. 1–11
31 Matthew, R.B., Jiebo, L.: 'Beyond pixels: Exploiting camera metadata for photo classification', *Patt. Recogn.*, 2005, **38**, (6), pp. 935–946
32 Farinella, G.M., Battiato, S.: 'Representation models and machine learning techniques for scene classification', in Wang, P.S.P. (Ed.): 'Pattern recognition and machine vision' (River publisher, 2010)
33 Torralba, A., Oliva, A.: 'Statistics of natural image categories', *Netw. Comput. Neural Syst.*, 2003, **14**, pp. 391–412
34 Torralba, A., Oliva, A.: 'Semantic organization of scenes using discriminant structural templates'. Int. Conf. on Computer Vision, 1999, pp. 1253–1258
35 Torralba, A., Oliva, A.: 'Depth estimation from image structure', *IEEE Trans. Patt. Anal. Mach. Intell.*, 2002, **24**, (9), pp. 1226–1238
36 Torralba, A., Pawan, S.: 'Statistical context priming for object detection'. Int. Conf. on Computer Vision, 2001, pp. 763–770
37 Webb, A.R.: 'Statistical pattern recognition' (Wiley, 2002, 2nd edn.)
38 Luo, J., Boutell, M.R.: 'Natural scene classification using overcomplete ICA', *Patt. Recogn.*, 2005, **38**, (10), pp. 1507–1519
39 Torralba, A.: 'Contextual priming for object detection', *Int. J. Comput. Vis.*, 2003, **53**, (2), pp. 169–191
40 Bhattacharyya, A.: 'On a measure of divergence between two statistical populations defined by probability distributions', *Bull. Calcutta Math. Soc.*, 1943, **35**, pp. 99–109
41 Comaniciu, D., Ramesh, V., Meer, P.: 'Kernel-based object tracking', *IEEE Trans. Patt. Anal. Mach. Intell.*, 2003, **25**, (5), pp. 564–575

doi: 10.1049/iet-cvi.2010.0056