Ridgelet-based signatures for natural image classification

Herve Le Borgne and Noel O'Connor

Center for Digital Video Processing, Dublin City University, Dublin 9,Ireland
hlborgne@eeng.dcu.ie, oconnorn@eeng.dcu.ie

RÉSUMÉ. *Dans cet article nous proposons une nouvelle représentation des images naturelles permettant de les organiser en groupes sémantiquement consistants. Les catégories concernées par la méthode sont identifiées par les propriétés statistiques des scènes naturelles. Les images sont décrites par une signature basée sur les* ridgelets. *Elle est combinée à une classifieur à vecteur support (SVM),qui est particulièrement adapté à la représentation des données en grande dimension, résultant en un système de reconnaissance efficace. Le potentiel de notre approche est démontré par une série de classifications binaires (e.g. ville/paysages or scènes extérieures/intérieures) sur une base de 1900 images.*

ABSTRACT. *This paper presents an approach to grouping natural scenes into (semantically) meaningful categories. The proposed approach exploits the statistics of natural scenes to define relevant image categories. A ridgelet-based signature is used to represent images. This signature is used by a support vector classifier that is well designed to support high dimensional features, resulting in an effective recognition system. As an illustration of the potential of the approach several experiments of binary classifications (e.g. city/landscape or indoor/outdoor) are conducted on databases of natural scenes.*

MOTS-CLÉS : *ridgelets, catégorisation, classification, médias non textuels.*

KEYWORDS: *ridgelets, categorisation, classification, non-textual medias.*

1. Introduction

For the last fifteen years, several fields of research have converged in order to address the management of multimedia databases. A new discipline has been created from this collective effort, usually called *Content-Based Image Retrieval (CBIR)* [SAN 01]. One of the key-issues, termed the *semantic gap*, is the lack of coincidence between the information extracted from the visual data at the lowest level (pixel values) and the interpretation that the same data has for a user in a given situation [SME 00]. Finding a general solution to this problem is a long-term research challenge that will involve not only computer science (integrated databases) and pictorial analysis but also cognitive sciences. Some successes have already been reported for particular problems, using various image processing and machine learning techniques. A comprehensive study of these works can be found in [SME 00]. It is well worth noting that the efficiency of a given technique is generally dependent on the particular image domain (or application) it was developed for. For instance, research efforts were specifically developed for face detection, in order to deal with the high degree of variability in appearance [Hje 01].

Among the wide variety of applications related to CBIR, this paper deals specifically with the management of *natural scenes*. This consists of the natural environment in the every-day life of a digital camera owner for instance. In fact, natural scenes have a particular statistical structure that has been widely studied in the literature, and we argue these properties can be used to address the semantic gap for this kind of content. In the context of CBIR, automatic image categorisation can help to large-scale image database retrieval and browsing by hierarchically classifying images into narrower categories that reduce the search time [VAI 98]. Such an organization may also be useful for image enhancement by allowing a selection of scene to which can apply a specific color and exposure adjustment.

In CBIR, one can broadly distinguish between the description of images and the computation of the similarity between images, even if both are closely linked. This paper mainly focus on determining which features must be extracted from natural images in order to group them in semantically meaningful categories. The usual approach to the classification of natural scenes is to use some general tools of image analysis that are justified *a posteriori* by their relative efficiency for a given problem. One of the first attempts in this direction was [GOR 94] in which the authors extracted the dominant direction of texture by a multiscale steerable pyramid to separate pictures of cities and suburbs from others. In [SZU 98], 1324 images were classified into indoor and outdoor classes using color (histogram), texture (autoregressive model) and frequency (discrete cosine transform) information. In [VAI 98] several two-class discriminations are performed using color histogram, color coherence vector, DCT coefficients, edge histogram and edge direction coherence vector. In [OLI 99], 700 images are classified into landscapes and “artificial scenes” that contains man-made structures such as buildings or roads. The authors employed a combination of Gabor filters to define prototypical “templates” of each category. A major contribution to the definition of content-based image descriptors was realized during the development of

the ISO/MPEG-7 standard. We refer to [MAN 01] for a comprehensive presentation of the standard and the descriptors. These descriptors were shown to be efficient for particular applications, but they do not take into account the statistical structure of natural images. On the contrary, we propose in this paper to fully exploit this structure to overcome the semantic gap for natural scenes. The representation of images is based on the ridgelet transform [CAN 99] that is optimally designed to represent edges in natural images.
The paper is structured as follows. In section 2, we discuss the statistics of natural scenes to motivate our approach. In particular, we explain how their power spectrum relates to the semantics of images categories. In section 3 we present the ridgelet transform and the representation model we propose for natural scenes. In section 4 experimental results show the efficiency of our representation using a Support Vector classifier. We conclude and present directions for future work in section 5.

2. Structure of natural scenes

A fundamental characteristic of natural images is their high degree of redundancy. If natural images were composed of random pixels, the redundancy would be null. They have a non-random statistical structure, resulting from regularities in the textures, shapes and surfaces of objects and scenes represented. Thus it follows that if we know what some sections of an image look like we are able to guess the missing sections [RUD 94]. Hence, among the set of all images, natural scenes form a particular subset that has been studied by many authors [SIM 01]. Because of the high dimensionality of the image space this subset is probably impossible to fully characterize, but some properties can be identified.

One of the most noticeable properties states that the average power spectrum of natural scenes decreases according to a law $1/f^{\alpha}$, where f is the spatial frequency and α is approximatively 2 (or 1 if one considers the amplitude spectrum instead of the power spectrum) [RUD 94]. As a first approximation, this was considered true regardless of the direction in the spectrum. Nonetheless, some studies refined this assertion [OLI 99, TOR 03]. Natural scenes with small depth ("closed scenes") have actually a spectrum in $1/f^2$ in all directions, but when the depth of the scene increases, the presence of a strong horizontal line enhances vertical frequencies (termed "open scenes"). Moreover, images representing human constructions contain a lot of horizontal and vertical lines, enhancing the corresponding frequencies. These characteristics are illustrated in figure 1. In [TOR 03] it was shown that some categories can then be defined according the shape of their spectrum, corresponding on the one hand to an approximate depth of the scene and on the other hand to a level of semantic meaning. It well worth noting the categories are defined by their global statistics, but that useful classification information can be extracted at local scales [TOR 03].
Rather than "yet an other global descriptor" to recognize any category of image, we propose to identify generic classes of scenes that are defined by their intrinsic properties. They form coherent perceptive sets of images that are congruent with a semantic

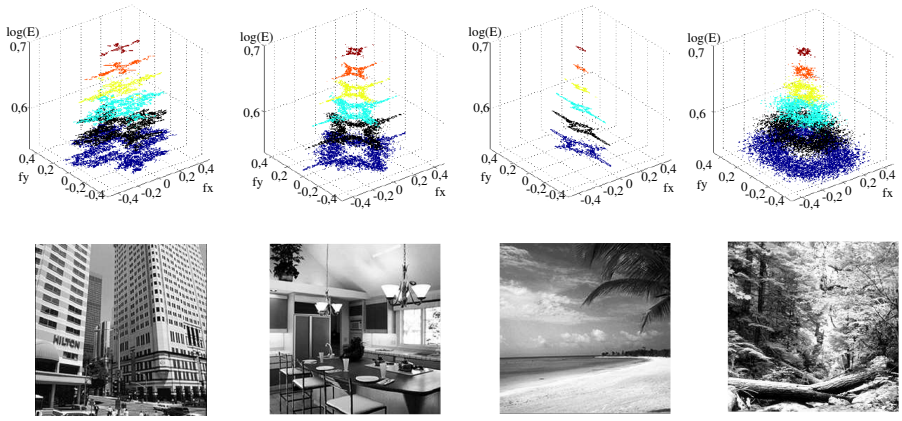

Figure 1. *Logarithm of prototypical power spectrum of natural scene categories. From left to right, categories are "outdoor cities", "indoor scenes", "open scenes" and "closed scenes" (from [OLI 99]).*

meaning. In a CBIR system, this can be used as a pre-classification step to organize large databases of natural images according to the user perception. Several tools can potentially serve as image descriptors, but it makes sense to choose one that fits to the properties of natural scenes, such as ridgelets.

3. Image representation

3.1. *Ridgelet*

The continuous ridgelet transform (CRT) of an integrable bivariate function $f(x)$ is defined by the following real function [CAN 99] :

$$CRT_f = \int_{\mathbb{R}^2} \psi_{a,b,\theta}(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} \tag{1}$$

where the bidimensional ridgelets $\psi_{a,b,\theta}(x)$ are defined from a unidimensional wavelet $\psi(x)$ as :

$$\psi_{a,b,\theta}(\mathbf{x}) = a^{-1/2}\psi\left(\frac{x_1 cos\theta + x_2 sin\theta - b}{a}\right) \tag{2}$$

where a is a scale parameter, b a shift parameter, and $\mathbf{x} = (x_1, x_2)^T$. Hence, a ridgelet is constant along the line $x_1 cos\theta + x_2 sin\theta = const$ and has the shape of the wavelet $\psi(x)$ in the perpendicular direction.
Finding a discrete form of the ridgelet transform is a challenging issue. The key point for this is to consider the CRT of an image as the 1-D wavelet transform of the slices (i.e. projections) of its Radon transform. This later is a simpler version of the Hough

transform that enhances the lines present in an image. An implementation of discrete ridgelets was proposed in [DO 02], based on the finite Radon transform. This method uses finite geometry and can then be applied to image of size $p \times p$ with p a prime number, which limits its use in our context. The method developed in [AVE 01] is based on the pseudopolar Fourier transform that evaluates the 2-D Fourier transform on a non-Cartesian grid. This transform is used to compute the Radon transform, and supports several nice properties, such as invertibility, algebraic exactness, geometric fidelity and rapid computation for images of size $2^n \times 2^n$. In the following we use the Beamlab package [DON 03] that implements the method proposed in [AVE 01].

3.2. *Image signature*

The ridgelet transform of an image corresponds to the activity of a mother ridgelet at different orientations, scales and spatial localizations. At a given orientation, there are 2^n localization at the highest scale, 2^{n-1} at the immediate lower scale, and so on until the lowest scale. For an image of size $2^n \times 2^n$, this results in a response of size $2^{n+1} \times 2^{n+1}$, introducing a redundancy factor equal to four.
In this work, we consider the global activity of ridgelets on the image because, as explained in section 2, it can fully characterize the semantic of the scene. As a consequence, we average the activity of similar ridgelets at different spatial locations. Hence we obtain one coefficient for each of the 2^{n+1} orientations and $n-1$ scales, resulting in a signature of size $(n-1) * 2^{n+1}$. Moreover, since the sign of the activity simply corresponds to an opposite contrast, we compute the average of the absolute value of the activity.
This scheme is valid for image of size $2^n \times 2^n$ only. In order to deal with images of any size, we extract the largest square part of the image and reduce it to an image of suitable size (i.e. the side is a power of two). This process determines the size of the signature that has a direct influence on the computation time and the recognition performance (see section 4.1). The scheme requires the central part of the image to be significant with respect to the whole scene. This assumption is very reasonable for images that are slightly rectangular (e.g. $4/3$ format) but may seem quite weak for particular formats such as $16/9$ or panoramic. If the statistics of the scene are stationary, any square part is representative and can be used to characterize the image. In the other case, the picture can not strictly be considered as a coherent scene and is out of the scope of this work. In this case, one could attempt to divide the picture in two or three parts with stationary statistics, and characterize each of them independently.

3.3. *Support vector classifier (SVC)*

To demonstrate the advantages of the proposed signature for natural scene recognition we use a support vector machines (SVM) [VAP 95] classifier, because of its efficiency to classify high dimensional data. The SVC is commonly used because of several attractive features, such as simplicity of implementation, few free parameters

required to be tuned, ability to deal with high-dimensional input data and good generalisation performances on many pattern recognition problems. This last property is due to the fact that SVM tend to minimise an upper bound on the expected risk (structural risk minimisation), while other learning techniques such as neural networks usually tend to minimise the error on the training set (empirical risk minimisation).
We now describe how SVM can be applied to classification in linear separable case. Let us consider a set of training samples $\{(x_i, y_i)_{1 \leq i \leq N}, x_i \in \mathcal{X}, y_i \in \mathcal{Y}\}$, with $\mathcal{X}$ the input space (e.g. $\mathbb{R}^D$), and $\mathcal{Y} \triangleq \{-1, +1\}$ the label space. In the linear case, one assumes the existence of a separating hyperplane between the two classes, i.e. a function $h(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$ parameterized by $(\boldsymbol{w}, b)$, such that the sign of this function applied to x_i gives its label. If such a function exists, then an infinity can be identified. By fixing $\min_i |h(x_i)| = 1$, we choose the normal vector $\boldsymbol{w}$ such that the distance from the closest point of the learning set to the hyperplane is $1/\|w\|$.

When the training data is not linearly separable, a more complex function can be used to describe the boundary. This is done by using a non-linear mapping of data into a potentially much higher dimensional feature space, in which a simple classification is easier to find [VAP 95]. In the following we will use a polynomial kernel of the form $(< xi, x_j > +1)^d$ where d varies from 0 to 3. The best result is then used to measure performance. Similar results were obtained when the value of d is fixed to 3.

Several methods were proposed to extent SVM to multi-class case [HSU 02]. The one-against-all method consists of constructing as many SVC as classes. The ith SVC is trained with all the example in the ith class with positive label and all other examples with negative labels. It results in a set of decision functions which the cardinal is the number of classes. Then a test data x_i is attributed to the class with the largest decision value. Another popular method is called the one-against-one method. For a set of C classes, it consists of constructing $C(C-1)/2$ binary classifiers where each is trained on data from two classes. Test data x_i is then classified by the C SVC, each of them giving a vote for the class to choose. The x_i is predicted to be in the class with the largest vote [FRI 96]. However, no theoretical result exists to determine the optimal method. Hence, in this paper we limit our approach to binary classifications, keeping the possibility of using one of the above strategies for multi-class classification in future works.

4. Experimental results

4.1. *Size of the signature*

The size of the signature depends directly on the size of the central square image on which the ridgelet transform is computed. The computation time of the signature mainly depends on this size and, less significantly on the original size of the image (because of the dimension reduction). Table 1 gives the size of the signature and its average computation time on 1903 images for different sizes of central square. The

Size of central square (pixel)	Length of signature	Computation time (seconds)
128×128	1536	1.66 ± 0.1
64×64	640	0.75 ± 0.05
32×32	256	0.40 ± 0.04
16×16	96	0.25 ± 0.03

Tableau 1. *Characteristics of the signature. Computation time is computed on 1903 images (3 Ghz CPU with 512 MB of RAM).*

interesting property of the proposed representation is that it can be easily adapted to time or storage capacity constraints.

4.2. *Recognition performances*

The database consists of 1903 images of different sizes collected on the web and professional databases[1]. They were divided into four classes : 452 *cities*, 405 *open* (landscapes with a marked horizon line), 544 *indoor* scenes and 502 *closed* scenes (landscapes without depth). Labels of the images were chosen according to the pre-classification of the professional databases and the judgement of two experts (one is the first author). The signatures of images are computed as explained in section 3.2. The classifier was implemented using the LibSVM package [CHA 01] with a polynomial kernel (see section 3.3). Two kinds of experiments were conducted : indoor *versus* outdoor (cities, open, closed) and cities *versus* landscapes (open, closed) with individual experiments focusing on classification between specific subcategories in each case. The binary classifications were repeated 20 times with randomly chosen learning and testing databases without overlap (cross-validation). The size of the learning database was fixed to 20 images, but larger sizes gave similar results on our databases.

In Table 2 we compare the performance as a function of the size of the signature. As expected, the recognition rate is an increasing function of size for all experiments. With the largest signature, classification rates are more than 87% for most experiments. The lower performance of city/indoor experiments is due to the similarity of their power spectrums (figure 1).
We also compare our signature to other descriptors. *Edge histograms* (EH) detects edges oriented at four directions ($0\,°$, $45\,°$, $90\,°$, $135\,°$) and one "non direction" in 16 adjacent regions of the image, resulting in a 80-dimensional vector. *Homogeneous texture* (HT) is computed by filtering the image with a bank of Gabor filters at 5 scales and 6 orientations. The descriptors contains the mean intensity, the standard deviation and the mean and standard deviation energy of the output filters, resulting in a 62-dimensional vector [MAN 01]. Contrary to the proposed signature that takes into

1. www.corel.com - www.goodshoot.com

Experiment	$Ridg_{128}$	$Ridg_{64}$	$Ridg_{32}$	$Ridg_{16}$
city/open	$93.0(2.0)$	$90.7(2.6)$	$86.2(1.5)$	$80.4(1.9)$
city/closed	$88.5(2.9)$	$84.3(2.6)$	$75.1(3.3)$	$65.2(3.4)$
city/indoor	$66.6(2.5)$	$63.4(2.8)$	$61.6(3.6)$	$56.7(2.0)$
open/indoor	$93.0(1.5)$	$89.7(1.5)$	$85.8(1.3)$	$80.2(2.4)$
closed/indoor	$87.5(2.2)$	$84.7(2.4)$	$75.7(2.9)$	$63.9(2.6)$

Tableau 2. *Average classification rate (average and standard deviation) for different size of signature on several experiments (20 cross-validations).*

account the central part of the image only, these two descriptors (EH and HT) are computed on the whole image. Results of classification, are given in Table 3. Our method is significantly better than the two others to discriminate cities from landscapes with at least 4% better classification rate than EH and 10% than HT. In the "indoor versus outdoor" paradigm, it outperforms Homogeneous Texture but is comparable to Edge Histograms for two experiments (taking into account the uncertainty measured by standard deviation) and lower for the open/indoor experiment.

Experiment	$Ridg_{128}$	**EH**	**HT**
city/open	$93.0(2.0)$	$88.8(1.7)$	$81.0(1.6)$
city/closed	$88.5(2.9)$	$82.3(2.2)$	$72.4(2.6)$
city/indoor	$66.6(2.5)$	$68.2(2.8)$	$63.3(2.2)$
open/indoor	$93.0(1.5)$	$96.7(0.6)$	$89.5(1.3)$
closed/indoor	$87.5(2.2)$	$88.6(1.6)$	$75.9(2.4)$

Tableau 3. *Comparison of our signature with* edge histogram *(EH) and* homogeneous texture *(HT).*

5. Conclusion and future work

In this paper, we proposed a new representation of natural images, based on a ridgelet description, which takes into account the statistical structure of natural image categories. These intrinsic statistical properties allow to identify perceptively coherent categories, for which the perception of the scenes match the semantic a human can attribute to the images. Hence, the method we proposed relevant and efficient to classify such images into the corresponding categories. In a CBIR system, this can be used as a pre-classification step to organize large databases of natural images according to the user perception.

In association with a support vector classifier, this results in an efficient scheme to discriminate cities from landscapes. We showed the limit of the current implementation of our method that is less efficient to separate pictures of outdoor cities from those of indoor rooms, but is still efficient to discriminate outdoor landscapes from indoor scenes.

Future work will deal with reduction of redundancy in the signature, in order to obtain a more compact representation of images and work on the implementation of the ridgelets, in order to obtain a faster computation. For instance identifying the most useful parts of the signature would allow to reduce the computation time of the signature by limiting the number of Radon projections required. This work is required to make our method applicable in a real case (e.g. 100,000 images) since the current computation cost is still too large. Finally, our method will be extended to the local statistical analyze, that was shown relevant to predict the presence of object (or people) in natural scenes.

6. Aknowledgement

We are grateful to the three anonymous referees who extensively reviewed our paper and pointed out several shortcomings, thus helping improve its quality. We thank Jovanka Malobabic for stimulating discussions at the early stages of this work. This research was supported in part by the European Commission under contract FP6-001765 aceMedia (URL : http ://www.acemedia.org) and by Enterprise Ireland through the Ulysses project number FR/2005/56.

7. Bibliographie

[AVE 01] AVERBUCH A., COIFMAN R. R., DONOHO D. L., ISRAELI M., WALDN J., « Fast slant stack : A notion of Radon transform for data in a Cartesian grid which is rapidly computable, algebraically exact, geometrically faithful and invertible », *SIAM Scientific Computing*, , 2001.

[CAN 99] CANDÈS E., DONOHO D., « Ridgelets : the key to high-dimensional intermittency ? », *Phil. Trans. Royal Society of London A*, vol. 357, 1999, p. 2495-2509.

[CHA 01] CHANG C., LIN C., « LIBSVM : a library for support vector machines », 2001, Software available at `www.csie.ntu.edu.tw/~cjlin/libsvm`.

[DO 02] DO M. N., VETTERLI M., « The finite ridgelet transform for image representation, », *IEEE trans. on Image Processing*, vol. 12, n° 1, 2002, p. 16-28.

[DON 03] DONOHO D., FLESIA A., HUO X., LEVI O., CHOI S., SHI D., « Beamlab 2.0 », website, january 2003, http ://www-stat.stanford.edu/ beamlab/.

[FRI 96] FRIEDMAN J., « Another approach to polychotomous classification », Technical report, 1996, Department of statistics, Stanford university, available at http ://www-stat.stanford.edu/ jhf/ftp/poly.ps.Z.

[GOR 94] GORKANI M., PICARD R., « Texture Orientation For Sorting Photos “At A Glance” », *ICPR-A*, vol. 1, 1994, p. 459-464.

[Hje 01] HJELMÅS E., LOW B., « Face Detection : A survey », *Computer Vision and Image Understanding*, vol. 83, 2001, p. 236-274.

[HSU 02] HSU C.-W., LIN C.-J., « A comparison on methods for multi-class support vector machines », *IEEE trans. on Neural Networks*, vol. 13, 2002, p. 415-425.

[MAN 01] MANJUNATH B., OHM J.-R., VASUDEVAN V., YAMADA A., « Color and texture descriptors », *IEEE trans. circuits and systems for video technology*, vol. 11, n° 6, 2001, p. 703-715.

[OLI 99] OLIVA A., TORRALBA A., GUÉRIN-DUGUÉ A., HÉRAULT J., « Global semantic classification of scenes using power spectrum templates », *Challenge of Image Retrieval*, Springer-Verlag, 1999, Newcastle, UK.

[RUD 94] RUDERMAN D., « The statistics of natural images », *Natwork : computation in neural systems*, vol. 5, 1994, p. 517-548.

[SAN 01] SANTINI S., *Exploratory image databases : content-based retrieval*, Academic press, London, 2001.

[SIM 01] SIMONCELLI E., OLSHAUSEN B., « Natural image statistics and neural representation », *Annual review of neurosciences*, vol. 24, 2001, p. 1193-1216.

[SME 00] SMEULDERS A. W. M., WORRING M., SANTINI S., GUPTA A., JAIN R., « Content-Based Image Retrieval at the End of the Early Years », *IEEE trans. on Pattern Analysis and Machine Intelligence*, vol. 22, n° 12, 2000, p. 1349-1380.

[SZU 98] SZUMMER M., PICARD R., « Indoor-outdoor image classification », *IEEE international workshop on content-based access of images and video databases*,, 1998, Bombay, India.

[TOR 03] TORRALBA A., OLIVA A., « Statistics of Natural Images Categories », *Network : Computation in Neural Systems*, vol. 14, 2003, p. 391-412.

[VAI 98] VAILAYA A., JAIN A., ZHANG H.-J., « On Image Classification : City Images vs. Landscapes », *Pattern Recognition*, vol. 31, n° 12, 1998, p. 1921-1936.

[VAP 95] VAPNIK V., *The Nature of Statistical Learning Theory*, NY :Springer-Verlag, 1995.