Natural scene classification and retrieval using Ridgelet-based Image Signatures

Hervé Le Borgne and Noel O'Connor

Centre for Digital Video Processing, Dublin City University, Dublin 9,Ireland
[hlborgne,oconnor]@eeng.dcu.ie

Abstract. This paper deals with knowledge extraction from visual data for content-based image retrieval of natural scenes. Images are analysed using a ridgelet transform that enhances information at different scales, orientations and spatial localizations. The main contribution of this work is to propose a method that reduces the size and the redundancy of this ridgelet representation, by defining both global and local signatures that are specifically designed for semantic classification and content-based retrieval. An effective recognition system can be built when these descriptors are used in conjunction with a support vector machine (SVM). Classification and retrieval experiments are conducted on natural scenes, to demonstrate the effectiveness of the approach.

1 Introduction and related works

For the last 15 years, several fields of research have converged in order to address the management of multimedia databases, creating a new discipline usually called *Content-Based Image Retrieval (CBIR)* [1]. One of the key-issues to be addressed, termed the *semantic gap*, is the disparity between the information extracted from the raw visual data (pixel) and a user's interpretation of that same data in a given retrieval scenario [2]. Automatic image categorization can help to address this issue by hierarchically classifying images into narrower categories, thereby reducing search time. Some successes have been reported for particular problems, using various image processing and machine learning techniques. In [3], the dominant direction of texture, estimated via a multiscale steerable pyramid allows identification of pictures of cities and suburbs. In [4], *indoor/outdoor* classification was achieved using color (histogram), texture (MSAR) and frequency (DCT) information. In [5], the authors hierarchically discriminate *indoor* from *outdoor*, *city* from *landscape*, and *sunset* from *forest* and *mountain* using color histograms, color coherence vectors, DCT coefficients, edge histograms and edge direction coherence vectors. However, none of these approach take into account the particular statistical structure of a natural scene, although this has been widely studied in the literature. One of the most noticeable properties states that the average power spectrum of natural scenes decreases according to $1/f^{\alpha}$, where f is the spatial frequency and α is approximatively 2 [6]. As a first approximation, this was considered true regardless of direction in the spectrum. Nonetheless, some studies have refined this assertion [7, 8]. Natural scenes with small perceived depth, termed *closed* scenes, do have a spectrum of $1/f^2$ in all directions, but when the depth of the scene increases, the presence of a strong horizontal line corresponding to the horizon enhances vertical frequencies. The latter type of images are termed *open* scenes. Moreover, images representing human constructions, termed *artificial*, contain a lot of horizontal and vertical lines and this reflected in the corresponding frequencies.

In [8], it was shown that some image categories can be defined, corresponding to an approximate depth of the scene (congruent to semantic), according the shape of their global and local spectrums. These properties were first used to address the semantic gap in [7], by classifying *landscapes* and *artificial scenes* using Gabor filters. In a similar vein, we exploit this statistical structure to address the semantic gap for natural scenes, using the ridgelet transform that is optimally designed to represent edges [9]. The main contribution of this work is to propose a method that reduces the size and the redundancy of this ridgelet representation, by defining both global and local signatures that are specifically designed for semantic classification and content-based retrieval. Section 2 presents the ridgelet transform and the associated proposed global and local signatures. Experimental results for image classification and retrieval using these descriptors are presented in section 3, with conclusions drawn in section 4.

2 Image representation

2.1 Ridgelet transform

Given an integrable bivariate function $f(x)$, its continuous ridgelet transform (CRT) is defined as [9]:

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

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 of its Radon transform. We used the method developed in [10], 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 support several nice properties, such as invertibility, algebraic exactness, geometric fidelity and rapid computation for images of size $2^n \times 2^n$. Code for this is provided in the Beamlab package [11].

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 localizations at the highest scale, 2^{n-1} at the next lowest scale, and so on. For an image of size $2^n \times 2^n$, this results in a response of size $2^{n+1} \times 2^{n+1}$. The challenge is therefore to create a signature for the image from these responses, that leads to a reduction of the size of the feature whilst preserving relevant information useful for discrimination.

2.2 Global ridgelet signature

The global ridgelet signature (Rd_{glb}) is extracted by averaging the ridgelet responses over all spatial locations. This is motivated by the reported possibility of defining semantic categories of natural scenes according to their global statistics [8]. Since ridgelets are computed on square images, we extract the largest square part of the image and reduce it to an image of size $2^n \times 2^n$. Keeping one coefficient for each of the 2^{n+1} orientations and $n-1$ scales results in a signature of size $(n-1) * 2^{n+1}$. Since the sign of the activity simply corresponds to contrast direction, the average of the absolute value of the activity is computed.

2.3 Local ridgelet signature

For this descriptor, the image is divided into $4 \times 4 = 16$ non-overlapping areas, and the ridgelet transform of each area is computed. Because of the same constraints as for the global signature, each area has actually a size $2^n \times 2^n$ pixels. It has been shown that narrower categories can be defined by such local statistics [8]. A local template is designed to compute the signature for each area. It defines 10 regions on which a measure of activity is computed, as shown on figure 1. Other local templates were designed but can not be presented in this paper due to space constraints. There are two regions at the lower frequencies and four at the middle and higher frequencies centered around $0\,°$, $45\,°$, $90\,°$ and $135\,°$. For each region, we compute the activity as the average absolute value of the ridgelet response divided by the standard deviation over the region. This gives 10 coefficients for each local signature i.e. 160 coefficients for the local ridgelet signature (Rd_{loc}).

2.4 Support vector classifier (SVC)

Support vector classifiers (SVC) [12] are commonly used because of several attractive features, such as simplicity of implementation, few free parameters required to be tuned, the ability to deal with high-dimensional input data and good generalisation performance on many pattern recognition problems.

To apply a support vector machine (SVM) to classification in a linear separable case, we consider a set of training samples $\{(x_i, y_i),\ x_i \in \mathcal{X},\ y_i \in \mathcal{Y}\}$, with $\mathcal{X}$ the input space, and $\mathcal{Y} \triangleq \{-1, +1\}$ the label space. In the linear case, we assume the existence of a separating hyperplane between the two classes, i.e

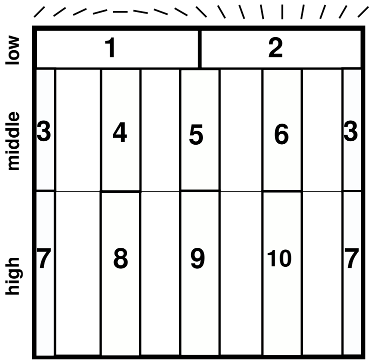

Fig. 1. Template defining 10 regions for the local signature. Rows represents the scales and columns are the orientations.

a function $h(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$ parameterized by $(\boldsymbol{w}, b)$, such that the sign of this function when applied to x_i gives its label. By fixing $\min_i |h(x_i)| = 1$, we chose the normal vector $\boldsymbol{w}$ such that the distance from the closest point of the learning set to the hyperplane is $1/\|w\|$. When training data is not linearly separable, a more complex function can be used to describe the boundary. This is done by using a kernel to map non-linear data into a much higher dimensional feature space, in which a simple classification is easier to find. In the following we use the LibSVM implementation [13] with a polynomial kernel of degree 1 to 4.

3 Experimental results

In this section the performance of our ridgelet signatures for image classification and retrieval are compared to that of descriptors defined in the MPEG-7 visual standard [14]: *Edge histograms* (EH), *Homogeneous texture* (HT) based on a Gabor filter description, *Color Layout* (CL) and *Scalable Color.*

3.1 Scene classification

Our test corpus consists of 1420 images of different sizes collected from both the web and professional databases[1] and are divided into four classes: *cities*, *indoor*, *open* and *closed* scenes. As explained in section 1, *open/closed* scenes refer to images of natural scenes with large/small perceived depth (i.e. with/without a horizon). Image signatures were computed as explained in section 2.2 and three sets of experiments were performed (Table 1). First, each class was classified against the others (Exps. $N°1 \ldots 4$). Then, *artificial* (consisting of both *cities, indoor*) versus *natural* (consisting of both *open, closed*) discrimination was investigated (Exp. $N°5$) as well as the intra-class classification within these classes (Exps. $N°6$, $N°7$). The final set of experiments investigated *cities* versus *natural*

[1] www.corel.com - www.goodshot.com

classification and the associated intra-class classification (Exps. $N°8 \ldots 10$). All experiments were repeated 10 times with randomly chosen learning and testing databases without overlap (cross-validation). The size of the learning database was fixed to 40 images, but larger sizes gave similar results.

Experimental results are presented in Table 1. To discriminate one class from the others, Rd_{loc} performs best for *cities* and *open*, while EH is better for *indoor* and *closed*. Color descriptors (CL and SC) have the worst performance, except for *indoor* for which all results are quite close, confirming the results of [4] that illustrated the importance of color descriptors for indoor/outdoor discrimination. In the *artificial* versus *natural* experiment, EH perform best, though Rd_{loc} has significantly better results than any other descriptor in the intra-class experiments ($N°6$, 7). EH and Rd_{loc} have similar results for *natural* versus *cities* classification ($N°8$). EH is slightly better in experiment $N°9$ but Rd_{loc} outperforms all others in discriminating *open* scenes from *cities* ($N°10$).

3.2 Scene retrieval

In order to estimate the retrieval performance of our signatures, the test corpus was extended to 1952 images using images from five smaller categories. Each of these new categories are characterized by the presence of an object: *door, firework, flower, car, sailing*. Objects are presented in a scene context that is congruent with their semantic: fireworks are in the sky, sailing activities on the sea, door on a building, cars on a road and flowers in a *closed* natural scene.

In practice, images are sorted according to their distance from the hyperplane calculated by the SVC. Retrieval performances are usually estimated by the probability of detecting an image given that is relevant (*recall*) and the probability that an image is relevant given that it is detected by the algorithm (*precision*). However, precision generally decreases according to the number of images detected while recall increases. Thus, precision is a function of recall ($p(r)$) and a trade-off must be chosen. The average value of $p(r)$ over $[0 \ldots 1]$ defines the *average precision* and measures retrieval performance taking into account both recall and precision.

Retrieval experiments were repeated ten times with different learning databases, and the average performance (measured by *average precision*) are shown in Table 2. For scenes with global characteristics (first four rows), the ridgelet signature performs well for *open* and *cities* but less so for *closed* and *indoor*. In this latter case, results for Rd_{loc} are similar to that of the color descriptors. Retrieval experiments for scenes containing a specific object (last five rows) demonstrate that Rd_{loc} is among the best results for three categories (*car, firework, sailing*) but has poor results for *flowers* and *doors*. In this latter case, Rd_{glb} has quite good performance though still significantly poorer than that of EH.

4 Concluding remarks

In this paper, we proposed a new representation of natural images, based on a ridgelet description. Two image signatures were designed, allowing both global and local analysis. When used in conjunction with a support vector machine, these descriptors can be used to classify natural scene categories. The proposed descriptors also exhibit good performance in retrieval of scenes containing specific categories of objects. Future work will focus on defining other local signatures to address the shortcomings of this approach for specific categories, and combination between global and local approach.

5 Acknowledgements

Part of this work was supported by the European Commission under contract FP6-001765 aceMedia (URL: http://www.acemedia.org). Authors are also supported by Enterprise Ireland and Egide that fund the Ulysse research project ReSEND (FR/2005/56).

References

1. Santini, S.: Exploratory image databases : content-based retrieval. Academic press, London (2001)
2. 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 **22** (2000) 1349–1380
3. Gorkani, M., Picard, R.: Texture orientation for sorting photos "at a glance". ICPR-A **1** (1994) 459–464
4. Szummer, M., Picard, R.: Indoor-outdoor image classification. In: IEEE international workshop on content-based access of images and video databases,. (1998) Bombay, India.
5. Vailaya, A., Jain, A., Zhang, H.J.: On image classification: City images vs. landscapes. Pattern Recognition **31** (1998) 1921–1936
6. Ruderman, D.: The statistics of natural images. Network: computation in neural systems **5** (1994) 517–548
7. Oliva, A., Torralba, A., Guérin-Dugué, A., Hérault, J.: Global semantic classification of scenes using power spectrum templates. In: Challenge of Image Retrieval, Springer-Verlag (1999) Newcastle, UK.
8. Torralba, A., Oliva, A.: Statistics of natural images categories. Network: Computation in Neural Systems **14** (2003) 391–412
9. Candès, E., Donoho, D.: Ridgelets: the key to high-dimensional intermittency? Phil. Trans. Royal Society of London A **357** (1999) 2495–2509
10. 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)
11. Donoho, D., Flesia, A., Huo, X., Levi, O., Choi, S., Shi, D.: Beamlab 2.0. website (2003) http://www-stat.stanford.edu/ beamlab/.
12. Vapnik, V.: The Nature of Statistical Learning Theory. NY:Springer-Verlag (1995)
13. Chang, C., Lin, C.: LIBSVM: a library for support vector machines. (2001) Software available at `www.csie.ntu.edu.tw/~cjlin/libsvm`.
14. Manjunath, B., Ohm, J.R., Vasudevan, V., Yamada, A.: Color and texture descriptors. IEEE trans. circuits and systems for video technology **11** (2001) 703–715

Table 1. Percentage of correct classification for our signature and MPEG-7 descriptors (see part 3 for notations). Results are the average (± standard deviation) classification rate for 10 cross-validations.

Exp. N°	**Experiment**	$\mathbf{Rd}_{glb}$	$\mathbf{Rd}_{loc}$	**EH**	**HT**	**CL**	**SC**
1	Cities *Vs* other	64.5(±3.9)	**77.2**(±4.2)	67.0(±3.4)	63.3(±4.5)	59.5(±2.6)	62.1(±4.0)
2	Closed *Vs* other	60.4(±4.2)	55.6(±1.8)	**73.6**(±3.1)	59.6(±3.7)	62.5(±5.1)	62.9(±5.0)
3	Indoor *Vs* other	71.1(±2.7)	74.6(±2.1)	**79.4**(±2.2)	72.2(±3.3)	**79.6**(±2.2)	75.6(±3.4)
4	Open *Vs* other	75.9(±3.6)	**93.0**(±1.3)	78.2(±3.2)	73.6(±3.9)	65.4(±4.0)	66.0(±3.1)
5	Artificial *Vs* Natural	74.6(±1.7)	72.8(±1.6)	**82.5**(±1.4)	71.6(±2.8)	64.3(±3.2)	67.7(±3.3)
6	Indoor *Vs* Cities	69.1(±2.0)	**88.1**(±1.4)	71.9(±2.0)	69.4(±3.2)	83.3(±2.7)	74.8(±2.2)
7	Open *Vs* Closed	71.0(±2.6)	**91.8**(±0.9)	72.9(±3.0)	66.8(±2.4)	65.3(±2.0)	61.3(±3.8)
8	Natural *Vs* Cities	73.2(±1.7)	**78.3**(±2.3)	**79.0**(±2.1)	69.4(±2.5)	57.5(±3.5)	63.6(±4.3)
9	Open *Vs* Cities	83.7(±2.2)	**96.5**(±1.0)	85.3(±1.5)	76.5(±2.7)	60.0(±2.7)	66.5(±1.9)
10	Closed *Vs* Cities	70.1(±2.2)	74.6(±1.3)	**78.2**(±1.8)	70.5(±3.3)	66.8(±1.9)	68.9(±2.3)

Table 2. Average precision for our signature and MPEG-7 descriptors (see part 3 for notations). It shows the average (± standard deviation) over 10 repetitions with different learning databases.

Class	**Size**	$\mathbf{Rd}_{glb}$	$\mathbf{Rd}_{loc}$	**EH**	**HT**	**CL**	**SC**
city	322	35.1(±3.7)	**48.1**(±4.8)	32.8(±3.4)	25.1(±4.2)	25.7(±2.9)	30.4(±3.4)
closed	355	19.6(±1.3)	24.3(±2.5)	**37.8**(±6.0)	25.5(±2.8)	32.3(±3.2)	31.2(±2.7)
indoor	404	36.6(±2.9)	42.2(±4.5)	**52.4**(±4.0)	38.5(±5.3)	41.1(±2.3)	41.6(±5.5)
open	339	46.6(±5.4)	**66.8**(±3.9)	54.7(±4.9)	39.6(±5.4)	34.9(±2.8)	30.5(±2.6)
car	136	18.2(±3.2)	**32.4**(±4.6)	29.5(±7.0)	10.7(±2.9)	12.1(±1.6)	10.4(±1.7)
doors	100	43.9(±6.0)	14.4(±4.2)	**74.8**(±5.9)	19.6(±6.9)	16.3(±3.5)	10.7(±2.5)
firework	100	41.7(±6.6)	62.7(±5.7)	**68.2**(±9.0)	55.0(±19.1)	36.7(±13.2)	34.3(±8.9)
flower	100	11.6(±2.8)	07.9(±2.4)	21.7(±6.6)	17.9(±7.4)	10.2(±3.7)	**28.6**(±11.2)
sailing	100	08.7(±1.3)	21.7(±3.1)	19.1(±4.1)	6.4(±2.4)	14.9(±5.3)	**24.2**(±8.1)