AN APPLICATION OF GRAPH COMMUTE TIMES TO IMAGE INDEXING

Régis Behmo[1,2]*, Nikos Paragios*[1]*, Véronique Prinet*[2]

[1] MAS, Ecole Centrale Paris, Grande Voie des Vignes,
F-92295 Châtenay-Malabry Cedex, France
[2] NLPR/LIAMA, Institute of Automation, Chinese Academy of Sciences,
P.O Box 2728, Beijing 100080, China

1. INTRODUCTION

In the field of satellite imaging interpretation, the means for a human agent to access and to process the available acquired data are not able to cope with the breadth and the quality of the data itself. This situation is paradoxical because it means we, as a scientific community, actually receive too much information to value it according to its true worth. The bottleneck that we face is the representation of the visual content of the satellite images. An automatic method for reliably describing the content of image subregions would allow us to index the image databases and to perform content-based queries on them. This, in turn, would open the door to precise automatic statistical measures and would therefore expand our large-scale analysis capability.

We introduce a resolutely new image representation that takes into account both the local appearance of regions as well as their relative layout. It is based on the measure of spectral properties of a graph built on a sparse set of interest points sampled in the image. These properties represent the distances between groups of interest points, where distance is computed in terms of similarity and spatial proximity. The relative importance of the appearance and the layout in the representation can be defined by two parameters; we observe that the bag-of-visterms [1], which dismisses all spatial information from the image representation, is a particular case of our approach. The idea of using attributed graph to represent image content has been introduced before [2]. However the approach of [2] is based on pixels groups and as such cannot really be applied to high resolution data.

Our representation was designed with the specific goal of content-based image retrieval in mind: regions that display similar but not exactly identical features and layouts should nonetheless have close representations. On the other hand, the information contained in the representation should be sufficiently rich to be able to discriminate between a large variety of visual classes. As a matter of fact, our approach is able to address both intra and inter class variability.

2. METHODOLOGY

The construction of our image representation proceeds in several steps. **First**, we sample interest points from the image. The choice of the detector/descriptor pair is arbitrary and can be made depending on the application and the type of visual data. **Second**, we build the *features graph* of the image: it is an unoriented weighted graph in which each interest point is a node and the nodes that are likely to belong to the same visual part are all the more strongly connected. We consider that interest points that belong to the same visual parts have close spatial positions and similar descriptors. Therefore we decide to connect each node i to its M nearest neighbours according to the distance $\Delta(i,j) = \Delta_{desc}(i,j)^{\alpha}\Delta_{geo}(i,j)^{1-\alpha}$. The relative contribution of the appearance and the spatial proximity can be weighted by $\alpha \in [0,1]$. Changing the value of M determines the connectivity of the features graph and the typical scale of the queried object subparts that the graph structure will capture.

Third, this features graph is "normalised" by grouping together the nodes that are assigned to the same codebook entries. The resulting graph is the *collapsed graph* in which each node represents a codebook entry and the weight of the edge between two nodes k, k' is equal to the sum of the weights of the edges that join nodes from codebook entries k and k' in the features graph. The collapsed graph is a structure that can be used to compare different images, contrary to the features graph. The matrix of distances between graph nodes is an appropriate choice for representing the collapsed graph, given an appropriate definition of how this distance is computed. We could simply use the transition matrix of the graph or the matrix of shortest paths between graph nodes. However, in problems where the presence or the accuracy of graph nodes is uncertain, as it is the case here, the shortest path distance lacks robustness and does not provide any statistical information about the structure of the graph. In this respect the notion of *commute times* between graph nodes is preferable. Considering a random walk on the nodes of the collapsed graph started at node k with a transition probability proportional to the edge weights, the commute time

$CT(k,k')$ between graph nodes k, k' is defined as the average number of steps required to reach k' for the first time and then to come back to k (see [3], [4] for details). Notice that commute times can take infinite values when the graph is not connected. It has been shown ([4], see [5] for a summary) that the matrix of commute times can be computed as a function of the eigenvectors and eigenvalues of the Laplacian of the graph. In fact, this same function can also be viewed as an embedding of the nodes of the graph in a space of dimension arbitrarily low, if we consider only the dominant eigenvalues of the diagonalisation of the Laplacian. Our image representation χ is a normalisation of the $K \times K$ commute time matrix, where K is the size of our codebook: $\chi(k,k') = exp\left(\frac{-CT(k,k')}{K}\right)$. Notice that for $M = 0$ the only non-zero terms are the diagonal elements that correspond to quantised features located in the image and χ is equal to the binary bag-of-visterms.

The obtained representation is very high-dimensional ($K(K+1)/2$ with K of the order of a few hundreds to a few thousands). We can reduce this dimensionality by considering each image as a node in a graph and by embedding the nodes of the graph in a space of low dimension, as mentioned above.

3. RESULTS

We compute this representation in order to perform image indexing tasks. The 512×512 images we index come from a database of 0.5m panchromatic Quickbird images of the area of Beijing, China. The training phase is mostly unsupervised: the training set is expected to be labelled, but not necessarily segmented. We build the representation of each image in our training and testing sets and reduce the dimensionality of the embedding. We can then train a linear SVM classifier to separate our classes.

We tested our approach with classes corresponding to different degrees of urbanisation and road networks. An experiment is illustrated on figure 1: we classify 126 images coming from two different classes "vegetation" and "road". We observe that embedding the image representations in a space of dimension 2 is already sufficient to obtain a good separation between the datasets. With an embedding dimension of 20 we obtain a $96,00\%$ good classification rate. These results are in fact closer to total recall since certain images could be labelled in both classes. More generally, the quality of the results demonstrate the validity and the relevance of our approach.

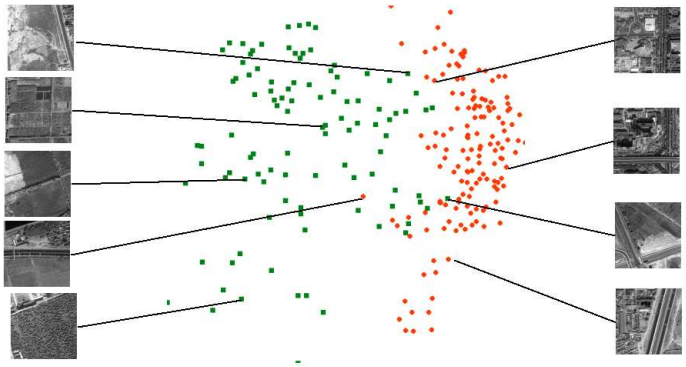

Fig. 1. "Vegetation" versus "Road" images classification.

4. REFERENCES

[1] P. Quelhas, F. Monay, J-M Odobez, D. Gatica-Perez, and T. Tuytelaars, "A thousand words in a scene," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 9, pp. 1575–1589, 2007.

[2] S. Aksoy, "Modeling of remote sensing image content using attributed relational graphs," in *SSPR/SPR*, 2006, pp. 475–483.

[3] F. R. K. Chung, *Spectral Graph Theory*, American Mathematical Society, February 1997.

[4] F. Chung and S. T. Yau, "Discrete green's functions," *Journal of Combinatorial Theory Series A*, vol. 91, no. 1-2, pp. 191–214, 2000.

[5] H. Qiu and E. R. Hancock, "Clustering and embedding using commute times," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 29, no. 11, pp. 1873–1890, 2007.