Automatic Classification of Outdoor Images by Region Matching

Oliver van Kaick and Greg Mori
School of Computing Science
Simon Fraser University, Burnaby, BC, V5A 1S6 Canada
E-mail: {ovankaic,mori}@cs.sfu.ca

Abstract

This paper presents a novel method for image classification. It differs from previous approaches by computing image similarity based on region matching. Firstly, the images to be classified are segmented into regions or partitioned into regular blocks. Next, low-level features are extracted from each segment or block, and the similarity between two images is computed as the cost of a pairwise matching of regions according to their related features. Experiments are performed to verify that the proposed approach improves the quality of image classification. In addition, unsupervised clustering results are presented to verify the efficacy of this image similarity measure.

1. Introduction

With the consolidation of the Internet as a medium for information interchange, several news agencies and educational organizations have allowed on-line access to their image and video collections in digital format. Moreover, with the popularization of digital cameras, more individuals are able to take pictures that can be shared in the Internet.

Therefore, a crucial task is to organize these large volumes of pictorial data, in order to extract relevant information. It is possible to address this problem by performing an automatic classification of images. This approach is characterized as classifying a number of images into different categories, where each category is composed of images that have similar content, in terms of high-level concepts.

The problem of image classification is specially challenging when considering outdoor images, which originate from a diversity of environments. Our goal is to distinguish amongst images according to relevant concepts, such as buildings, water, mountains, and vegetation.

This paper presents a method to compute the similarity between two images according to region matching. The method is applied to a dataset of outdoor images. In particular, we address the problems of image labelling – assigning labels to input images based on a hand-annotated set of training images – and unsupervised image clustering – grouping images into different categories according to their content. The proposed method consists in the following steps.

Firstly, the images to be classified are segmented according to a graph-based approach, which provides a good partition of the image into regions that represent relevant structures. We will compare this to a method that partitions the image into rectangular regions, providing a number of regular blocks.

Secondly, low-level features are extracted from the obtained regions, which can be either the largest irregular segments in the image or regular blocks. Next, the similarity between two images is computed as the cost of a pairwise matching of regions, taking into consideration the extracted features. For the problem of image clustering, the costs of region matching between each pair of images compose a similarity matrix, which is clustered to convey the requested number of image groups. For image labelling, a nearest neighbor classifier is used.

The proposed method differs from previous approaches by computing image similarity based on region matching. The cost of the best region assignment is used as a measure of similarity between two images. Therefore, it is possible to relate images that have similar content which appears in different spatial locations.

Thus, the main contribution of this work is the development of this novel region matching method, and showing that it improves the quality of automatic image classification. In addition, the difference between using regular blocks or segmented regions is also presented, which allows to verify that, for the proposed approach, general segmentation does not influence significantly the quality of the results.

This paper is organized as follows. Section 2 introduces some relevant previous work related to image classification based on low-level features. Moreover, the method proposed in this paper is described in Section 3. A number of experiments were performed to verify the validity of the proposed method. Details on how these experiments were conducted and the results that were obtained are showed in Section 4. Finally, Section 5 presents the conclusions for this work.

2. Previous Work

In order to group images into meaningful categories using only a set of low-level features, it is necessary to first extract a suitable set of features from the images, and then group the images into different categories according to these features. Therefore, Section 2.1 describes the different approaches proposed in the literature to address the problem of image classification. Section 2.2 describes specifically which features or sets of features are used in the different works to perform the intended classification.

2.1. Image classification

The most common approach proposed for the specific problem of content-based image retrieval is to store images in a database in conjunction with a set of features, that describe the main properties of these images. This approach led to the proposal of different content-based retrieval systems, such as QBIC [11], PhotoBook [13], and BlobWorld [3]. A survey presenting an overview of systems based on this approach and the problem of image retrieval from the World Wide Web is the work of Kherfi *et al.* [7].

However, these systems only allow to retrieve images that are similar to a query, not being able to classify images into meaningful groups. Due to the limited nature of these systems, parallel research concentrated in classifying images into meaningful categories, or relying in user feedback to further improve query results [15].

The first efforts following the idea of a general image classification propose to use a classifier to decide whether an image belongs to certain semantic class or not. These works include techniques that decide whether an image is indoor/outdoor [18, 16], or city/landscape [21].

Vailaya *et al.* [20] summarize these different works into one single idea, proposing a small but semantically meaningful binary hierarchy of *vacation* images, where the first level of the hierarchy is indoor/outdoor images, the second is city/landscape, and the other levels are related to specific classes of natural scenes. A Bayesian classifier is used to assign an image to its best related group, for each level of the binary hierarchy.

Further efforts propose methods for classification in each of these specific domains, such as outdoor [6] and indoor [8] environments, the latter being still a very studied topic due to its relevance to the field of robot localization and navigation.

Moreover, instead of using a hierarchical model, Oliva and Torralba [12] propose to use a holistic representation of a scene in terms of attributes that are intuitive for humans, such as naturalness, openness, roughness, expansion, and ruggedness of an image. These attributes are defined in terms of the overall characteristics of the Fourier decomposition of the image, and they provide information about the probable semantic category that the image belongs to.

More recent work takes a step further by trying to relate the problem of general image classification to the task of object location and recognition.

Li *et al.* [9] propose a generative/discriminative learning method, which is trained with a set of pre-annotated images. The method learns classes of objects that appear in an image and subsequently is able to automatically annotate a given outdoor scene. This approach also allows to combine different features that are extracted from an image, such as color, texture, and structure.

Furthermore, Torralba *et al.* [19] perform place recognition using a set of features combined with a hidden Markov model. This model indicates the likelihood of an image being related to a certain place, based on the given features and the last visited location. In their work, context priming is also used to predict the presence of certain objects in the scenes.

Moreover, a different approach is proposed by Duygulu *et al.* [4], which combines image classification and object location. An image is segmented into regions, which are are annotated according to a machine translation approach, by learning a mapping between these regions and annotated keywords. Therefore, the method predicts words for images as a whole, thus indirectly performing the task of image classification, and also assigns these words to specific regions of the image, implying object location.

In addition, for many of these described methods, one important aspect is which set of features is used to train classifiers or to provide general information about an image. The different features that are used in some of these works are presented in the next section.

2.2. Image Features

Most of the features that are successfully used in content-based retrieval systems are summarized in the work of Ma and Zhang [10]. The features presented in their work were originally employed for the task of content-based image retrieval or image classification by some of the approaches described below.

The work of Vailaya [20] *et al.* makes use of a number of features, performing image classification based on color moments and color histograms in HSV and Luv color-spaces, edge direction histograms, and color coherence vectors. Localization of features in their work is only performed for the color moments, which is accomplished by dividing each image into 10×10 blocks and computing the first and second order moments for each one of these 100 blocks.

Oliva and Torralba [12] use the energy spectrum of the Fourier Transform as the main set of features that describe an image. Different properties of this spectrum are used to characterize a scene, based on values obtained by a linear regression fitted to this data. Localization of features in this work is also accomplished by dividing each image into 8×8 blocks and computing a windowed Fourier transform for each one of these blocks.

Moreover, Renninger and Malik [14] advocate that scene recognition is performed initially by humans uniquely by means of texture analysis. Therefore, they propose a system where image classification is achieved by grouping images according to the responses of spatial filters, which are based on derivatives of Gaussians at different scales. This system achieves results similar to the ones obtained by humans, under the constraint of having to characterize an image in a short period of time.

Furthermore, wavelet filters, which also capture texture and edge information, are used in the work of Torralba *et al.* [19] for scene classification. Since filter responses usually are of the same or higher dimensionality of the filtered images, it is necessary to perform some quantization of the data. In their work, the average for large rectangular regions is simply computed to perform the required quantization.

Wang *et al.* [22] take a step further and introduce a more elaborated feature localization by segmenting an input image and extracting features from the obtained regions. The similarity between two images is estimated as a weighted sum of the similarity of the region features. This method is closely related to the one proposed in this work. However, the features used by the work of Wang *et al.* are more related to the characteristics of these regions, such as size, position and orientation, than to the their content, which is characterized mainly by color.

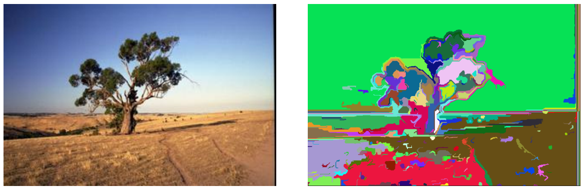
In the next section, the method proposed in this paper for image classification is described, in combination with the features that are used for region characterization.

3. Proposed Method

The purpose of the proposed method is to classify a given set of images into meaningful groups. Its basic idea is to establish the similarity between two images as being the cost of performing a pairwise matching of segmented regions. This pairwise matching is computed by taking into consideration features extracted from each region. Next, the matrix composed of the image similarities is clustered to convey the requested number of classes. The method is described in more detail as follows.

3.1. Segmentation

Firstly a rough segmentation of the images to be classified is performed. The graph-based segmentation method of Felzenszwalb and Huttenlocher [5] is used for this task. If the image is firstly widely blurred with a Gaussian filter, a segmentation consisting in a small number of large regions is obtained, which is suitable for the problem being addressed here. The result of such a segmentation for an example image is shown in Figure 1.



Input image

Segmented with [5]

Figure 1. Example of graph-based image segmentation.

Moreover, one question that naturally arises in this context is whether segmented regions are really necessary for the problem of image classification. Therefore, in this work, regular blocks are alternatively used instead of segmented regions. This approach consists in dividing an image into rectangular regions according to a certain factor. Regular blocks or segmented regions are equivalently used as regions to be matched.

3.2. Features

Furthermore, features are extracted from the larger segments or regular blocks that are obtained. These features are described as follows.

Color moments The first and second order moments, the mean and standard deviation, are extracted for each band of the HSV and Luv color-spaces. These color-spaces are preferred since HSV accounts for hue and saturation information, while Luv conveys more perceptually related color information.

Color histograms Histograms for the HSV and Luv color-spaces are also extracted.

Edge direction histograms Firstly, Canny edge detection is performed considering the whole input image. Furthermore, the orientation of a pixel that represents an edge is accumulated in a histogram, for the related block or region.

Texture The input image is firstly convolved with a set of 39 filter banks similar to the ones used by Renninger and Malik [14], consisting in oriented odd-symmetric filters, oriented even-symmetric filters, and radially symmetric center-surround filters, based on Gaussian kernels and computed at 6 different orientations and 3 scales.

Since the result of applying such filters are 39 images of the same dimension that the input images, the responses have to be quantized. This is achieved by simply computing the mean for 8 large portions of each block or region of the image, for each filter response.

The features that are extracted from each block or segment are used in the pairwise region matching described in the next section.

3.3. Region assignment

For each pair of images being considered, the method proceeds by computing the best pairwise matching between the regions of the two images. In order to accomplish this task, the Hungarian method that solves the weighted bipartite matching problem is used. Besides providing the best pairwise matching, the method also computes an associated cost, which is taken as a measure of similarity between the two images.

The matching cost is defined as

$$H = \min_{\pi} \sum_{i} C(p_i, q_{\pi(i)}) \tag{1}$$

where C denotes the similarity between the region p_i of the first image and the region q_j of the second image, and π is a permutation of the regions of the second image. The similarity C is computed as the distance (L_2 norm) between the feature vectors extracted from these two regions. Each obtained value is subtracted from the maximum cost, so that low cost values result in high similarity values.

3.4. Clustering

Furthermore, the matching cost values compose a similarity or affinity matrix, where each entry relates a pair of images. Spectral clustering is employed to cluster this matrix and obtain the requested number of image classes [17], which is a parameter specified by the user.

Since no semantic meaning is imposed to the clusters, it is not expected that all images in an obtained group represent a single concept. However, what is expected is that all the images that belong to a certain cluster are highly related, representing similar environments or having in common the presence of certain structures, such as buildings or trees.

3.5. Image labelling

Moreover, since it is difficult to come with an objective way of measuring the quality of the classes that are obtained, a variation of the described method is also investigated in this work. It consists in using a dataset of annotated images to verify the quality of the proposed method in terms of image labelling.

More specifically, the image dataset is divided into training and test images. Each test image is compared against all training images and the k nearest neighbors are selected, which are the images that possess the lower associated matching cost. Furthermore, the union of the annotations of the k training images is assigned to the test image and compared to the real annotations, which allows to compute precision and recall rates for specific keywords.

This procedure can also be modified to be suitable for a content-based retrieval system with more constrained requirements. For relating a query image to one of the obtained clusters, a set of representative images for each cluster is selected, and the query image is assigned to the cluster for which the region matching returns the lowest cost for its representative images. The representative images can be selected as being the ones with highest similarity when considering intra-cluster image matching.

The next section describes the details of the experiments that are conducted to verify the quality of the results obtained by the proposed method.

4. Experiments and Results

For the experiments, a dataset of annotated images from the University of Washington is used, which is entitled *Object and Concept Recognition for Content-Based Image Retrieval* [1]. This dataset was firstly presented in the work of Li *et al.* [9]. An example of the annotations related to two images is shown in Figure 2. The images have a resolution of 700×500.

It should be noted that the annotations are not perfectly coherent, since similar images tend to possess very different keywords. However, the annotations can be used to perform an objective analysis of the quality of the results, given by the number of matched keywords.

For feature extraction, a 3×3 median filter is firstly applied to the images, to eliminate extreme outliers. The color histograms, edge direction histograms, and color coherence vectors that are then extracted for each block or segment possess 64 bins each. For the initial image segmentation, the Gaussian filter that is used for blurring the input image has a parameter $\sigma = 3$.

The next two sections describe the results obtained by the proposed method for the clustering and the labelling problem, respectively.

beach ocean sky cliffs

water buildings boats trees sky

Figure 2. Examples of annotated images from the dataset used in the experiments.

4.1. Clustering

Figure 3 shows the complexity of the clustering problem, in terms of the similarity matrix that is obtained, and the resulting clustered matrix. In addition, Figure 4 shows some examples of the image clusters that are obtained with the proposed method, when the 25 largest regions obtained by the segmentation are considered for matching. A set of representative images is shown for each cluster.

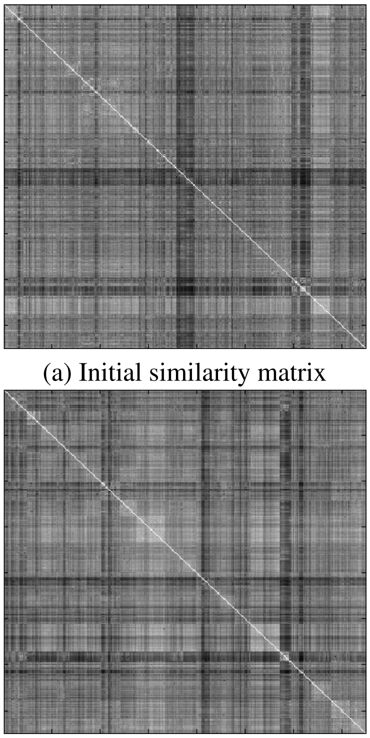
4.2. Labelling

When assessing the quality of the results according to the precision and recall of matched annotations, 374 images are used for training, while the remaining 505 images are used for testing. Each test image is compared against all training images, and the 7 nearest matchings are selected. A word is assigned to a test image if it is present in the annotations of at least 2 of the 7 nearest training images.

Furthermore, in order to evaluate the results more objectively, the annotations of the images are modified. Words which represent equivalent entities are substituted by only one word which represents the underlying common concept. For example, words such as *ocean*, *lake*, and *sea* are all substituted by the keyword *water*.

The total precision and recall obtained for all annotation keywords is presented in Table 1, as a consequence of the number of segments or regular blocks that are used for region matching. The first entry of the table presents the situation where images are compared without using regions, only according to global feature distance. The second entry shows the situation where no region matching is performed. The blocks are assigned directly according to their spatial location in the images, that is, the rectangle on the top-left corner of the first image is assigned to the rectangle on the top-left corner of the second image, and so on, for all other blocks.

It can be seen from these results that using region matching significantly improves the results in terms of recall, however, there is not much difference when segmented regions or regular blocks are used.



(a) Initial similarity matrix

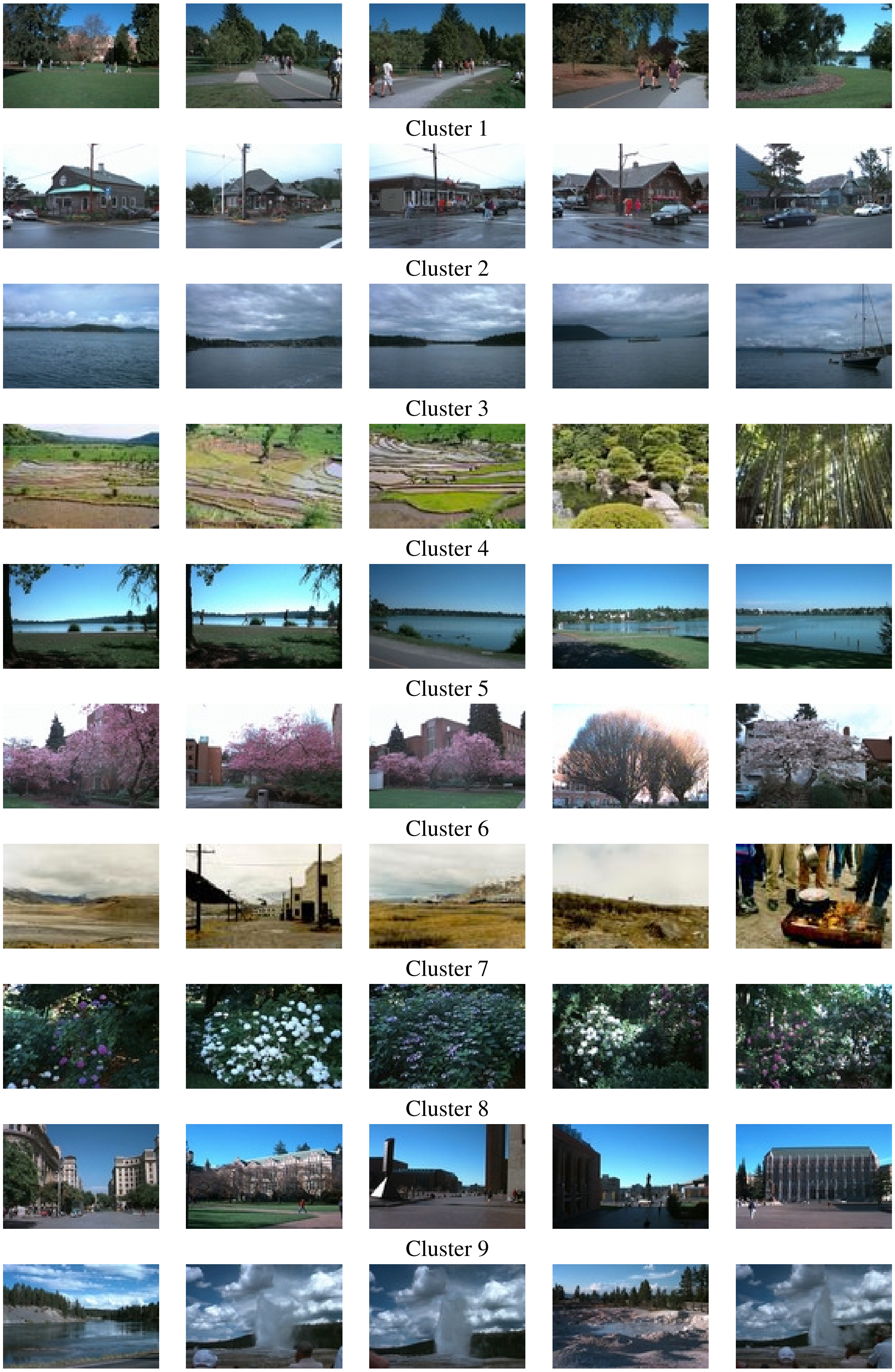
(b) Clustered similarity matrix

Figure 3. Results for the clustering of the similarity matrix. Lighter values indicate higher similarity.

Method	Precision	Recall
global	0.38	0.63
25 blocks*	0.42	0.56
4 blocks	0.46	0.70
9 blocks	0.46	0.69
25 blocks	0.46	0.70
4 segments	0.42	0.69
9 segments	0.42	0.69
25 segments	0.42	0.69

Table 1. Total precision and recall for different numbers of regular blocks and segments. (*) is the situation where no region matching is performed. For 25 blocks, the image is divided into 5×5 rectangles.

Moreover, Figure 5 shows the precision and recall obtained for each annotation keyword, when comparing the approach that uses 25 regular blocks to the one which uses



Cluster 1

Cluster 2

Cluster 3

Cluster 4

Cluster 5

Cluster 6

Cluster 7

Cluster 8

Cluster 9

Cluster 10

Figure 4. Examples of image clusters obtained with the proposed method.

25 segmented regions. Figure 6 presents the comparison between region matching and the global approach. Figures 5 and 6 also show graphs of the weighted harmonic mean of precision and recall, the F-measure defined as follows [2]

$$F(\text{recall}, \text{precision}) = 2 \times \frac{\text{recall} \times \text{precision}}{\text{recall} + \text{precision}} \qquad (2)$$

These measurements also imply that using segments or regular blocks for region matching are practically equivalent, when considering the precision and recall for individual annotation keywords. However, using region matching significantly improves the quality of the results.

One possible reason for the equivalence of performance of both types of regions might be that the obtained segmentation is not coherent for images with similar content. Objects that are present in one region of a given image may appear in separated regions for other images. Thus, the feature descriptors of each region change considerably, providing an inferior matching. Although segmentation should capture the structural content of a scene, unstable segmentation appears to be equivalent to regular blocks.

5. Conclusions

This paper presented a method directed towards the problem of automatic outdoor image classification. The proposed approach computes the similarity between two images as the cost of the best pairwise matching of regions. This considerably improves the quality of image labelling compared to a baseline method. Perhaps surprisingly, it can be concluded that, for this specific approach, the quality of the results does not differ significantly if regular blocks or segments are used as the regions for similarity computation.

References

[1] Object and concept recognition for content-based image retrieval (image database). `http://www.cs.washington.edu/research/imagedatabase/`.

[2] S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in images via a sparse, part-based representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(11):1475–1490, 2004.

[3] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Region-based image querying. In *Proc. IEEE Workshop on Content-Based Access of Image and Video Libraries*, pages 42–49, 1997.

[4] P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proc. European Conference on Computer Vision (ECCV)*, 2002.

[5] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *Int. J. Comput. Vision*, 59(2), 2004.

[6] G. L. Foresti. Outdoor scene classification by a neural tree-based approach. *Pattern Analysis & Applications*, 2(2):129–142, 1999.

[7] M. L. Kherfi, D. Ziou, and A. Bernardi. Image retrieval from the world wide web: Issues, techniques, and systems. *ACM Comput. Surv.*, 36(1):35–67, 2004.

[8] J. Kosecka, L. Zhou, P. Barber, and Z. Duric. Qualitative image based localization in indoors environments. In *Proc. Computer Vision and Pattern Recognition*, volume 2, pages 3–8, 2003.

[9] Y. Li, L. G. Shapiro, and J. A. Bilmes. A generative/discriminative learning algorithm for image classification. In *Proc. of IEEE Int. Conf. on Computer Vision (ICCV)*, volume 2, pages 1605–1612, 2005.

[10] W. Y. Ma and H. J. Zhang. Content-based image indexing and retrieval. In B. Furht, editor, *The Handbook of Multimedia Computing*, chapter 11. Boca Raton, 1998.

[11] W. Niblack, R. Barber, W. Equitz, M. Flickner, E. Glasman, D. Petkovic, P. Yanker, C. Faloutsos, and G. Taubin. The QBIC project: Querying images by content using color, texture and shape. In *Proc. SPIE Storage and Retrieval for Image and Video Databases*, volume 1908, pages 173–187, 1993.

[12] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vision*, 42(3):145–175, 2001.

[13] A. Pentland, R. W. Picard, and S. Sclaroff. Photobook: Content-based manipulation of image databases. *Int. J. Comput. Vision*, 18(3):233–254, 1996.

[14] L. W. Renninger and J. Malik. When is scene recognition just texture recognition? *Vision Research*, 44:2301–2311, 2004.

[15] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: A power tool for interactive content-based image retrieval. *IEEE Trans. Circuits Syst. Video Technol.*, 8:644–655, 1998.

[16] N. Serrano, A. Savakis, and J. Luo. A computationally efficient approach to indoor/outdoor scene classification. In *Proc. Int. Conf. on Pattern Recognition*, volume 4, pages 146–149, 2002.

[17] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

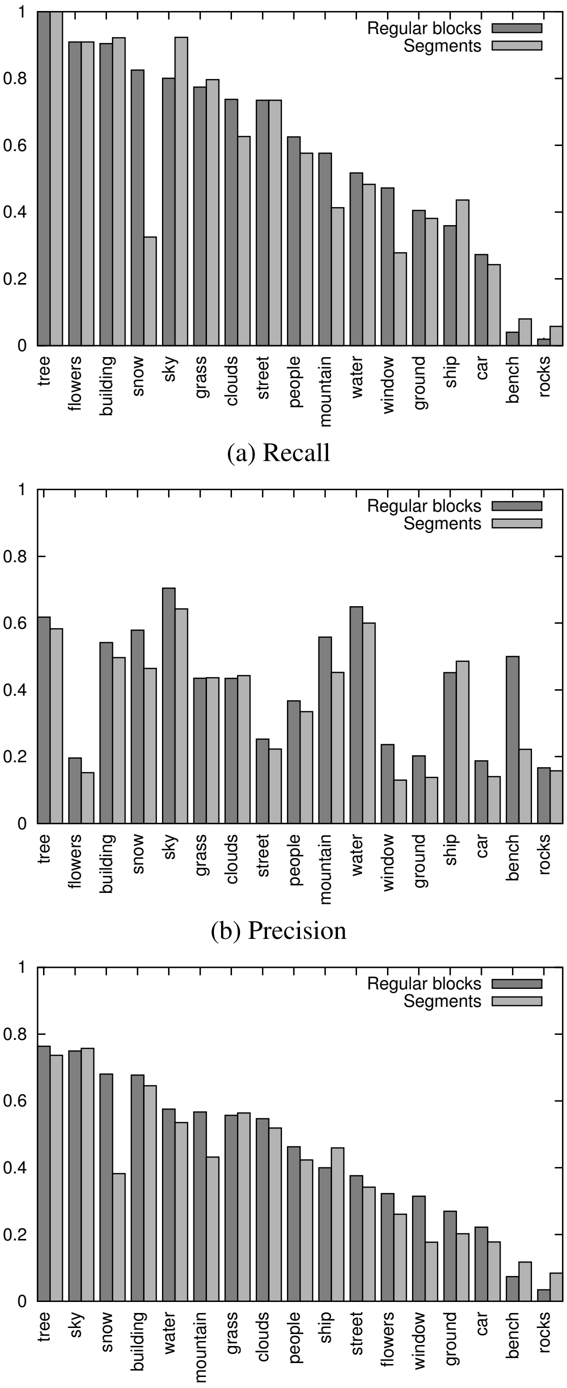
[18] M. Szummer and R. W. Picard. Indoor-outdoor image classification. In *Proc. of IEEE Int. Workshop on Content-Based Access of Image and Video Database*, pages 42–51, 1998.

[19] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin. Context-based vision system for place and object recognition. In *Proc. 9th IEEE Int. Conf. on Computer Vision*, volume 1, pages 273–280, 2003.

[20] A. Vailaya, M. A. T. Figueiredo, A. K. Jain, and H. J. Zhang. Image classification for content-based indexing. *IEEE Transactions on Image Processing*, 10(1), 2001.

[21] A. Vailaya, A. Jain, and H. J. Zhang. On image classification: City vs. landscape. In *CBAIVL '98: Proceedings of the IEEE Workshop on Content-Based Access of Image and Video Libraries*, pages 3–8, 1998.

[22] T. Wang, Y. Rui, and J.-G. Sun. Constraint based region matching for image retrieval. *Int. J. Comput. Vision*, 56(1–2):37–45, Jan. 2004.

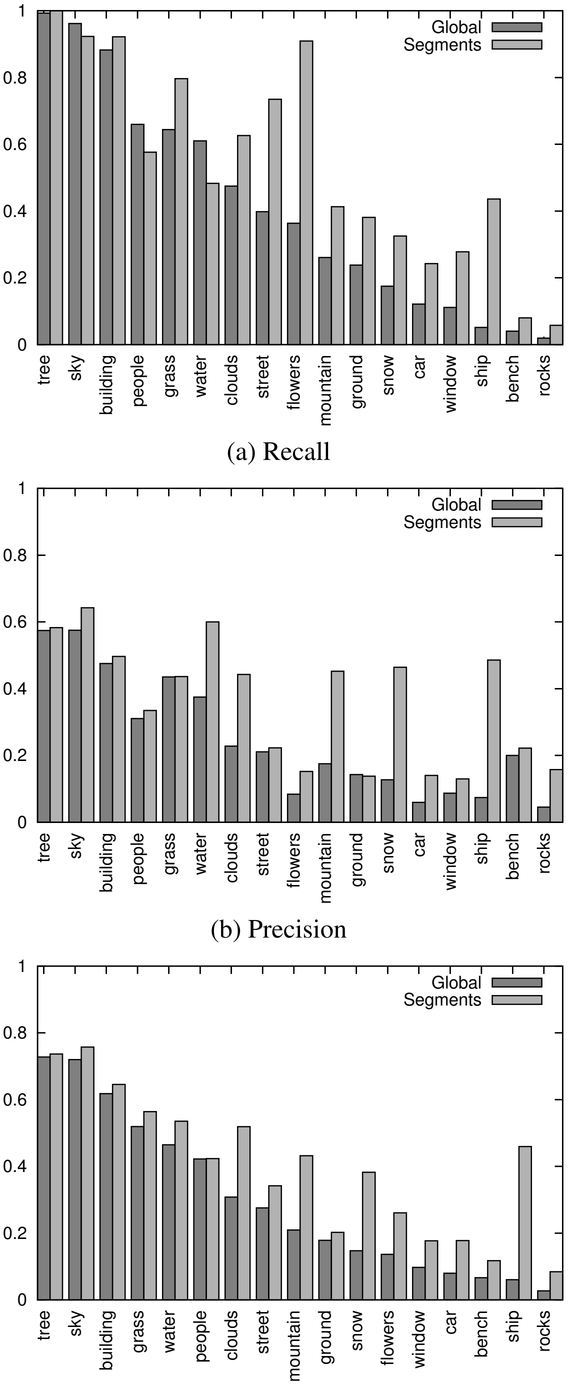


(a) Recall

(b) Precision

(c) F-measure combining recall and precision

Figure 5. Results obtained for region matching, when considering annotation keywords.



(a) Recall

(b) Precision

(c) F-measure combining recall and precision

Figure 6. Comparison between region matching and the global approach, when considering annotation keywords.