Natural Scene Image Modeling using Color and Texture Visterms.

Pedro Quelhas[1,2] and Jean-Marc Odobez[1,2]

[1] IDIAP Research Institute
[2] Ecole Polytechnique Federale de Lausanne (EPFL)

Abstract. This paper presents a novel approach for visual scene representation, combining the use of quantized color and texture local invariant features (referred to here as *visterms*) computed over interest point regions. In particular we investigate the different ways to fuse together local information from texture and color in order to provide a better *visterm* representation. We develop and test our methods on the task of image classification using a 6-class natural scene database. We perform classification based on the *bag-of-visterms* (BOV) representation (histogram of quantized local descriptors), extracted from both texture and color features. We investigate two different fusion approaches at the feature level: fusing local descriptors together and creating one representation of joint texture-color visterms, or concatenating the histogram representation of both color and texture, obtained independently from each local feature. On our classification task we show that the appropriate use of color improves the results w.r.t. a texture only representation.

1 Introduction

Viewpoint invariant local descriptors [1, 2] (i.e. features computed over automatically detected local areas) have proven to be useful in long-standing problems such as viewpoint-independent object recognition, wide baseline matching, and image retrieval. These feature were designed to have a high degree of invariance. As a result, they are robust to changes in viewpoint and lighting conditions. Furthermore, due to their locality, they provide robustness to image clutter, partial visibility, and occlusion. In addition, the use of *quantized* local invariant features has also proven in recent years to provide a robust and versatile way to model images, leading to good classification [3–5], retrieval [6, 7] and image segmentation [4, 8] performance.

A great advantage of modeling images based on quantized local invariant features for the tasks of retrieval and classification is that the same methodology can be used for different image categories and that performance is often similar if not better than most of the existing task specific state-of-the-art algorithms. This was demonstrated on images of objects [3] and on scenes [4, 5]. Moreover, in scene classification, this general approach performed surprisingly well given that only local texture features were used [4, 5], while most state-of-the-art techniques [9, 10] are based on both texture and color. Nevertheless, in visterm based representations used for scene classification, it seems that by discarding color we are potentially eliminating discriminative information since several scene classes are characterized by specific colors. Thus, it is quite natural and relevant to investigate the use of color in visterm based approaches and address the following related questions: how can color be integrated in the BOV framework and how much is gained by doing so?

In this paper, we propose and present an approach to model scene images using both color and texture visterm representations. More precisely, we first show on a 6-class problem that texture based invariant local features, used to build bags-of-visterm representations, are suitable for natural scene classification. Secondly we show that the inclusion of color improves the classification results. Although not demonstrated in the paper, the representation and methods presented here can be extended for ranking/retrieval [3–5].

The rest of the paper is organized as follows. The next Section discusses related work. Section 3 presents the BOV image representation. Section 4 describes the experimental setup. Classification results are provided and discussed in Section 5. Section 6 concludes the paper.

2 Related work

The problem of image modeling using low-level features has been studied in image and video retrieval for several years [9–13]. Broadly speaking, the existing methods differ by the definition of the target image classes, the specific image representations, and the classification method. In the next paragraphs, we focus our dicussion on the image representation issue.

Image representations based on quantized invariant local descriptors have been used for many tasks, with variations on both local detectors/descriptors and the subsequent image representation. Sivic et. al. [6] applied text retrieval methodologies on quantized local descriptors in a movie keyframe retrieval application. The system was proven to be fast and usable for large image database queries. The exploitation of quantized local descriptors was further extended by Csurka et. al. [3] to object recognition. The authors proposed to represent images using an histogram of the quantized local descriptors (bag-of-visterms). On the Caltech object image database, their system was show to outperform state-of-the-art object recognition techniques. In more recent work Quelhas et. al. [4] and Fei-Fei et. al. [5] have show that this bag-of-visterms representation can be further decomposed into mixtures of latent semantic models. Such latent models enable clustering and ranking of images into meaningfull groups.

On the field of scene image modeling, most works use color and texture information to perform classification/retrieval. Vailaya et al. [9] used histograms of different low-level cues to perform scene classification. Different sets of cues were used depending on the two-class problem at hand: global edge features were used for city vs landscape classification, while local color features were used in the indoor vs outdoor case. More generally scene recognition methods tend to fuse color and texture information. Both Serrano et al. [14] and Szummer et al. [10] propose a two-stage classification of indoor/outdoor scenes, where color and texture features of individual image blocks are computed over a spatial grid layout are first independently classified into indoor or outdoor. The local classifica-

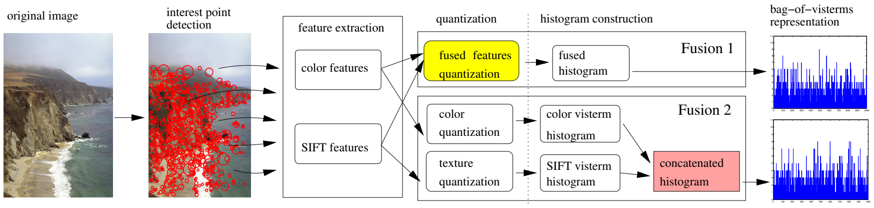

Fig. 1. Schematic representation of our system and of the two alternative fusion approaches: fusion between texture and color information is done either at the feature level before quantization (yellow box) or at the bag-of-visterm level(pink box).

tion outputs are then further combined to create the global scene representation used in the final image classification stage. Vogel and Schiele propose a similar two-stage approach based on a fixed grid layout to perform scene retrieval and classification [15]. Several local features (color and edge histograms, grey-level co-occurence matrix) are concatenated after normalization and weighting into one feature vector. Finally, Boutell et. al. [16] use only Luv color moments in a 7x7 block layout to perform scene multi-label scene classification.

In contrast, methods based on quantized local descriptors use only gray-scale information to create their fundamental features. Although this may be acceptable for some classes it is obvious that for natural scenes, color is important and its use may improve the power of the visterm representation.

3 Image modeling

In this section we first describe the bag-of-visterms (BOV) image modeling methodology. We then introduce the considered local features used in this paper, and finally introduce the considered fusion schemes.

3.1 Bag-of-visterms representation from local descriptors

The construction of the BOV feature vector h from an image d involves the steps illustrated in Fig. 1. In brief, interest points are automatically detected in the image, then local descriptors are computed over those regions. These descriptors are quantized into visterms, and the number of occurrences of each specific visterm of the vocabulary are counted to build the image BOV representation. In the following we describe in more detail each step involved in the construction of the BOV representation.

Interest point detection The goal of interest point detectors is to automatically extract characteristic points -and more generally regions- from the image, which are invariant to some geometric and photometric transformations. This invariance ensures that given an image and its transformed version, the same points will be extracted from both and hence, the same image representation will be obtained. Several interest point detectors exist in the literature. They vary mostly by the amount of invariance they theoretically ensure, the image property they exploit to achieve invariance, and the type of image structures they are designed to detect [2, 17].

In this work, we use the difference of Gaussians (DOG) point detector [2]. This detector essentially identifies blob-like regions where a maximum or minimum of intensity occurs in the image, it is invariant to translation, scale, rotation and constant illumination variations. This detector was selected for the following reasons. First, it was shown to perform well w.r.t. other detectors [17]. Second, since it defines regions that are homogeneous, it is more adapted to the computation of color descriptors (e.g. mean colors) than, for instance, edge corners (Harris detector). Finally, as an increase of the degree of invariance may remove information about the local image content that is valuable for classification, the DOG detector is preferable than fully affine-invariant ones [1, 18].

Local descriptors Local descriptors (SIFT or color moments, see next Subsection) are computed on the region around each interest point detected by the local interest point detector.

Quantization and Vocabulary model construction When applying the two preceeding steps to a given image, we obtain a set of real-valued local descriptors. We then quantize each local descriptor into one of a discrete set $\mathcal{V}$ of visterms v according to a nearest neighbor rule:

$$\mathbf{v} \longmapsto Q(\mathbf{v}) = v_i \Longleftrightarrow \mathrm{dist}_{\mathrm{Q}}(\mathbf{v}, v_i) \leq \mathrm{dist}_{\mathrm{Q}}(\mathbf{v}, v_j) \qquad \forall j \in \{1, \ldots, N_{\mathcal{V}}\} \tag{1}$$

where $N_{\mathcal{V}}$ denotes the size of the vocabulary (the set of all visterms). The vocabulary is constructed by applying the K-means algorithm to the set of local descriptors extracted from the training images, and keeping the means as visterms. Except in the fusion case (see Section 3.3), we used the Euclidean distance in the clustering. As for vocabulary size, we used 1000 clusters since it has been shown that little performance can be gained by increasing this number [4, 5] when using texture visterms. However, when building a joint color-texture vocabulary, 2000 clusters were considered as detailed in Section 3.3.

Bag-of-visterms representation Finally, the BOV representation of the image is constructed from the local descriptors according to:

$$h(d) = (h_i(d))_{i=1..N_{\mathcal{V}}}, \text{ with } h_i(d) = \mathrm{n}(d, v_i) \tag{2}$$

where $\mathrm{n}(d, v_i)$ denotes the number of occurrences of visterm v_i in image d. This vector-space representation of an image contains no information about spatial relationship between visterms.

3.2 Local descriptors

In this work, two local descriptors were considered: SIFT (Scale Invariant Feature Transform) [2], representing local texture/structure information, and Luv color moments. The choice of SIFT was motivated by the findings of several publications [6, 17, 5], where SIFT was found to work among the best on several tasks. For color, the use of the Luv color space was motivated by the fact that it is a perceptual color space (it was designed to linearize the perception of color distances) and that it has also been known to perform well in both retrieval and recognition applications [9, 16]. A description of both features follows:

- **SIFT descriptor $\mathbf{v}_s$:** this descriptor is based on the gray-scale image. SIFT features are local histograms of edge directions computed over different parts of the interest region. They capture the structure of local image regions, which correspond to specific geometric configurations of edges or to more texture-like content. In [2], it was shown that the use of 8 orientation directions and a 4×4 grid gives a good compromise between descriptor size and accuracy of representation. The final feature size is thus 128. Orientation invariance is achieved by estimating the dominant orientation of the local image patch and normalizing for rotation. For a more compact representation, we applied a principal component analysis (PCA) decomposition on this features using training data. By keeping 95% of the energy, we obtain a 44-dimensional feature vector. The PCA step did not increase or decrease performance, but allowed for faster clustering.
- **Luv descriptor $\mathbf{v}_c$:** it is based on 121 Luv values computed on a 11×11 grid normalized to cover the local area given by the interest point detector. From these values, we calculate the mean and standard deviation for each dimension and concatenate the result into a 6-dimensional vector. Each dimension of this vector are then normalized to unit variance so that L (luminance) does not dominate the distance metric.

3.3 Feature fusion

We investigated two fusion strategies, addressing two different aspects of early fusion [19].

Fusion 1: in this approach, we fused the real valued SIFT and color features before quantization, in order to obtain a joint color/texture vocabulary (see top of Fig 1). The fusion occurs by concatenating the sift feature $\mathbf{v}_s$ and color feature $\mathbf{v}_c$, after normalization, and weighted by a mixing value α according to:

$$\mathbf{v} = (\alpha \mathbf{v}_s^\star, (1-\alpha)\mathbf{v}_c^\star) \text{ with } \mathbf{v}_s^\star = \beta_s \mathbf{v}_s \text{ and } \mathbf{v}_c^\star = \beta_c \mathbf{v}_c \tag{3}$$

The normalization factor β_s (resp. β_c) is learned by setting it to the inverse of the average euclidian distance between 50000 random pairs of SIFT (resp. color) features. These values were found to be: β_s =60 and β_c =1.6. As a consequence of this concatenation, using a euclidian distance $\mathrm{dist}_{\mathrm{Q}}$ in the Kmeans algorithm, we end up with a weighted distance between the two feature type distances:

$$\mathrm{dist}_{\mathrm{Q}}(\mathbf{v}^1, \mathbf{v}^2) = \alpha\, \mathrm{dist}_{\mathrm{Q}}(\mathbf{v}_s^{\star,1}, \mathbf{v}_s^{\star,2}) + (1-\alpha)\, \mathrm{dist}_{\mathrm{Q}}(\mathbf{v}_c^{\star,1}, \mathbf{v}_c^{\star,2}) \tag{4}$$

where the distance between feature types is approximately of the same order of magnitude. The value of α is learned through cross-validation on training data.

Fusion 2: here (bottom part of Fig. 1), we assumed that the two feature types (SIFT, color) were independent, and fused the features by concatenating their BOV representation, again using a mixing value α, i.e. $h = (\alpha h_s, (1-\alpha) h_c)$. The use of the mixing value is necessary to allow the weighting of the different BOV representation in the SVM classifier (see next Section).

4 Experimental setup

In this section we describe the database we used, the protocol we followed, and the baseline system used for comparison.

4.1 Database

We use the database kindly provided to us by Vogel et. al. [15], and which is constituted of 6 different natural scenes type images. This data set contains a total of 700 images of resolution 720×480 pixels, distributed over the 6 natural scene classes as follows: coasts (142), river/lakes (111), forests (103), plains (131), mountains (179), and sky/clouds (34). We chose this data because of its good resolution and color. Additionally, it is the only available public database we found which contained several natural classes. A drawback, however, is that the database has some non negligeable overlap between classes (e.g. an image belonging to a given class could also easily belong to another class given its content). This is a property originally introduced as part of the database to evaluate human classification performance.

4.2 Classifier and evaluation protocol

In all our experiments, we used a multi-class Support Vector Machine (SVM) for classification. To perform experiments, we adopted a 10-fold training/testing protocol. That is, data is split into 10 folds, and for each fold, all parameters (quantization, mixing value α, SVM capacity) are trained on the remaining 90% data, and the learned system is tested on the given fold. The presented class performance corresponds to the averages over the 10 runs, and the overall system performance is the macro average of the class performance.

4.3 State-of-the-art baselines

Baseline of [15] We considered as first baseline the approach introduced along with the database [15]. In that work, the image was divided into a grid of 10×10 blocks, and on each block, a feature vector composed of a 84-bin HSI histogram, a 72-bin edge histogram, and a 24 features grey-level co-occorence matrix was computed. These features were concatenated after normalization and weighting, and used to classify (with an SVM) each block into one of 9 local semantic classes (water, sand, foliage, grass,...). In a second stage, the 9 dimensional vector containing the image occurence percentage of each regional concept was used as input to an SVM classifier to classify images into one of the 6 scene classes. The reported performance of that approach were good: 67,2%[3].

Color histogram In order to compare the results obtained with our color-only visterm BOV representation against a more traditional color histogram approach, we used a concatenated Luv 96-bins linear histogram (32 bins for each dimension: L, u, and v).

[3] Note however that the approach of [15] requires much more work than ours, as labeled data (image blocks with labels) to train the intermediate regional concept classifier are necessary. In [15], approx. 70000 blocks were manually labeled!

Class	confusion matrix						performance
coasts	61.3	9.9	1.4	9.2	17.6	0.7	61.3
river/lakes	18.0	30.6	9.9	12.6	24.3	4.5	30.6
forests	0.0	0.0	90.3	2.9	6.8	0.0	90.3
plains	15.3	11.5	6.1	55.7	7.6	3.8	55.7
mountains	10.1	6.1	2.8	6.1	73.7	1.1	73.7
sky/clouds	14.7	2.9	0.0	14.7	0.0	67.6	67.6
overall							63.2

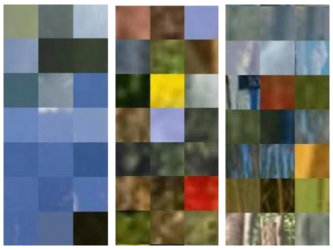

Table 1. Classification performance for the SIFT based BOV representation (left). Sample patches belonging to three visterms.

Class	confusion matrix						performance
coasts	49.3	16.9	2.8	12.7	15.5	2.8	49.3
river/lakes	21.6	31.5	9.0	7.2	30.6	0.0	31.5
forests	4.9	8.7	70.9	7.8	7.8	0.0	70.9
plains	9.2	9.2	6.9	53.4	16.8	4.6	53.4
mountains	12.3	12.3	1.7	14.0	59.2	0.6	59.2
sky/clouds	14.7	11.8	0.0	14.7	0.0	58.8	58.8
overall							53.9

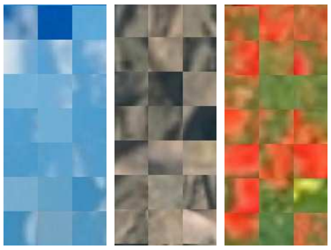

Table 2. Classification results for the Luv color space based BOV representation. Sample patches belonging to three random visterms.

5 Results

In this section, we first present the classification performance when using a single information source (texture or color), and then using the fusion schemes.

5.1 SIFT and color visterm BOV classification performance

Let us first explore the classification results obtained using the BOV representation constructed from each feature type separately.

SIFT features Table 1 provides the result obtained with the SIFT based BOV representation. While being slightly lower than the baseline, this approach performs surprisingly well given that no color information is used. This is illustrated by the sample patches belonging to 3 different visterms (Table 1, right). As can be seen (and as expected), visterm patches have no coherence in terms of color.

Color features Table 2 shows the results obtained using the Luv color visterms. Although significantly smaller than the performance obtain with SIFT visterms, the result is still relatively good given the features' simplicity (6 dimensions). Overall, all classes are affected by the performance degradation. Surprisingly, the forest class gets the most degradation, indicating that there is more reliable information in the local structure than in the color. When observing samples associated to some visterms (Table 2, right), we can see that the goal of color coherence is achieved, but that coherence in terms of texture/structure is mainly lost (there remain some coherence due to the specific interest point detector employed). To further analyse the performance of our BOV approach,

Class	confusion matrix						performance
coasts	69.0	8.5	2.1	7.7	10.6	2.1	69.0
river/lakes	21.6	28.8	9.0	11.7	26.1	2.7	28.8
forests	1.9	1.9	85.4	2.9	7.8	0.0	85.4
plains	9.2	9.2	2.3	62.6	12.2	4.6	62.6
mountains	8.4	5.6	1.1	5.6	77.7	1.7	77.7
sky/clouds	5.9	0.0	0.0	14.7	2.9	76.5	76.5
overall							66.7

Table 3. Classification results with the first fusion strategy: joint texture/color visterms. Sample patches belonging to three visterms.

we compared it with a simple Luv color histogram (see Section 4.3). The system performance in this latter case exhibited a strong performance drop, achieving a 34.1% recognition rate. This illustrates the necessity for both a data-driven and local approach, embedded in out BOV representation, as compared to global approaches based on more arbitrary color representations.

5.2 Fusion classification performance

We now present the classification results combining color and texture information in the BOV representation, as presented in Section 3.3.

Fusion 1: In this approach, local features are concatenated prior to clustering, resulting in a joint texture/color vocabulary of 2000 visterms. The average mixing value α obtained through cross-validation was 0.8, indicating that more importance was given to the SIFT feature. Table 3 displays the results obtained in this case. These results shows an overall improvement w.r.t. those based on the SIFT feature alone, and are very close to the baseline results (67.2%). The sky/clouds class is the one that beneficiate mostly from the improvement, with a reduction of its overlap with the coasts class.

When looking at the constructed vocabulary, we observe that visterms have coherence in both texture and color, as illustrated in Table 3. However, since now both features influence the clustering process, we notice an increase of the noise level in both color and texture coherence within the clusters.

Fusion 2: In this second strategy, it is assumed that, at the interest point level, information gathered from color is independent from texture/structure information. This strategy thus works by concatenating the BOV representation of color and texture, after having them weighted by the factor α. Interestingly enough, the optimal α value was again found to be 0.8, showing again an emphasis on information arising from the SIFT features. Table 4 shows the obtained results. These are nearly identical to those obtained with the first fusion strategy, and again very close to those of the baseline.

Overall, the results are encouraging, and demonstrate that the two approaches are valid for the scene classification task. Both fusion approaches performed significantly better than grey-scale BOV representation. The fact that both approaches reach similar results to the baseline may indicate that we are reaching the performance limit that may be obtained in this data when not using any

Class	confusion matrix						performance
coasts	58.5	13.4	1.4	13.4	10.6	2.8	58.5
river/lakes	20.7	36.0	7.2	9.9	23.4	2.7	36.0
forests	1.9	1.0	89.3	2.9	4.9	0.0	89.3
plains	12.2	6.1	6.9	64.1	7.6	3.1	64.1
mountains	6.1	7.3	3.4	6.7	76.0	0.6	76.0
sky/clouds	14.7	0.0	0.0	11.8	0.0	73.5	73.5
overall							66.2

Table 4. Classification results with the second fusion strategy: concatenation of the texture and color BOV representation.

spatial information. Figure 2 shows some images with the labels attributed by each systems we tested. We can see that some labels are subjective. For some images several possible labels could be considered correct. As such some of the errors that the systems produce seem logical. This indicates that the BOV representation captures valid scene properties, however this dataset does not supply a clear enough class definition for the training of the systems.

6 Conclusion

We investigated the use of color information, in addition to texture, to represent scene images with BOV relying on interest point detectors. Two fusion schemes were proposed and tested on a 6 class scene recognition task. They have shown that a small but significant gain in classification performance can be achieved w.r.t. texture only BOV representation. The obtained performances are similar to a state-of-the-art approach that requires much more supervised training. We believe that the proposed fusion schemes can easily be applied to other tasks.

Several extensions to the proposed BOV framework could be investigated to improve scene classification results. For instance, some invariance could be removed in the SIFT descriptor computation: recent studies in object recognition have shown that eliminating the rotation invariance usually leads to better results. Or, as the BOV discards all image spatial information, it would be interesting to reintroduce this information, for instance by computing regional BOV.

Acknowledgements

The authors acknowledge financial support by the Swiss National Center of Competence in Research (NCCR) on Interactive Multimodal Information Management (IM)2, and by the MULTImodal Interaction and MULTImedia DataMining (MULTI) project. Both projects are managed the Swiss National Science Foundation on behalf of the federal authorities.

References

1. Mikolajczyk, K., Schmid, C.: Scale and affine interest point detectors. International Journal of Computer Vision **60**(1) (2004) 63–86
2. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2) (2004) 91–110
3. Willamowski, J., Arregui, D., Csurka, G., Dance, C., Fan, L.: Categorizing nine visual classes using local appearance descriptors. In: Proc. of LAVS Workshop, in ICPR'04, Cambridge (2004)

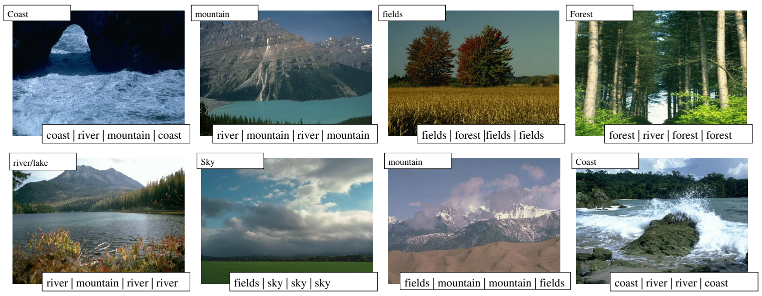

Fig. 2. Images ilustrating the resulting classification of the evaluated systems. Ground-truth is shown on the top left corner. On the bottom are all attributed labels: SIFT BOV, Color BOV, feature fusion and histogram fusion (from left to right).

4. Quelhas, P., Monay, F., Odobez, J.M., Gatica-Perez, D., Tuytelaars, T., Gool, L.V.: Modeling scenes with local descriptors and latent aspects. In: Proc. of IEEE Int. Conf. on Computer Vision, Beijing (2005)
5. Fei-Fei, L., Perona, P.: A Bayesian hierarchical model for learning natural scene categories. In: Proc .of. IEEE Int. Conf. on Computer Vision And Pattern Recognition, San Diego (2005)
6. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: Proc. of IEEE Int. Conf. on Computer Vision, Nice (2003)
7. Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T.: Discovering object categories in image collections. In: Proc. of IEEE Int. Conf. on Computer Vision, Beijing (2005)
8. Dorko, G., Schmid, C.: Selection of scale invariant parts for object class recognition. In: Proc. of IEEE Int. Conference on Computer Vision, Nice (2003)
9. Vailaya, A., Figueiredo, M., Jain, A., Zhang, H.: Image classification for content-based indexing. IEEE Trans. on Image Processing **10**(1) (2001) 117–130
10. Szummer, M., Picard, R.: Indoor-outdoor image classification. In: IEEE International Workshop CAIVD, in ICCV'98, Bombay (1998)
11. Oliva, A., Torralba, A., Guerin-Dugue, A., Herault, J.: Global semantic classification of scenes using power spectrum templates. In: Proc. of the Challenge of Image Retrieval, Newcastle upon Tyne, UK (1999)
12. Paek, S., S.-F., C.: A knowledge engineering approach for image classification based on probabilistic reasoning systems. In: Proc. of IEEE Int. Conference on Multimedia and Expo, New York (2000)
13. Smeulders, A., 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**(12) (2000) 1349–1380
14. Serrano, N., Savakis, A., Luo, J.: A computationally efficent approach to indoor/outdoor scene classification. In: Int. Conf. on Pattern Recognition. (2002)
15. Vogel, J., Schiele, B.: A semantic typicality measure for natural scene categorization. In: Pattern Recognition Symposium DAGM'04, Tübingen, Germany (2004)
16. Boutell, M., Luo, J., Shen, X., C.M.Brown: Learning multi-label scene classification. Pattern Recognition **37**(9) (2004) 1757–1771
17. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. In: Proc. of IEEE Int. Conf. on Comp. Vision and Pattern Recognition. (2003)
18. Matas, J., Chum, O., Martin, U., Pajdla, T.: Robust wide baseline stereo from maximally stable extremal regions. In: Proc. of the British Machine Vision Conference, Cardiff (2002)

19. Kittler, J., Hatef, M., Duin, R., Matas, J.: On combining classifiers. IEEE PAMI **20**(3) (1998) 226–239