Mixture Distributions for Weakly Supervised Classification in Remote Sensing Images

Jean-Baptiste Bordes[1]
[1] Telecom ParisTech,
F-75634 Paris Cedex 13, France

Véronique Prinet[2]
[2]NLPR/LIAMA/CAS
P.O Box 2728, Beijing 100190, China

Abstract

For its simplicity and efficiency, the bag-of-words representation based on appearance features is widely used in image and text classification. Its drawback is that shape patterns of the image are neglected. This paper presents a novel image classification approach using a bag-of-words representation of textons while taking into account spatial information. A generative probabilistic modeling of the distribution of textons is proposed. The parameters of the mixture's components are estimated using a EM algorithm. We show that the number of classes in a database can be found automatically and exactly by MDL. This modeling gives very good results for the task of weakly supervised classification in satellite images.

1 Introduction

Image classification is designed to assign one —out of several— class to each pixel or to each local-patch of a *visual scene*[1]. Automatic classification is relevant in a variety of applications ranging from text/image/video indexing and retrieval, to object categorisation, surveillance systems, medical and remote sensing images understanding [22, 12, 3, 4, 14]. In particular, in recent years, thanks to the advent of very high resolution (VHR) satellite imagery and its diffusion via web-based tools such as *GoogleEarth*, remote sensing (RS) applications attracted an increasing interest in the Machine Vision and Image Processing communities [3].

Textons, defined as the atomic visual elements of a visual scene, were first introduced by Julesz at the early stage of visual perception [11]. In his primal sketch representation [15], Marr extended this concept to image primitives, namely the "symbolic tokens", where primitives were defined as geometric features computed in a deterministic way. If the concept of the atomic structure endured, its definition has evolved: it is nowadays widely acknowledged that image primitives are features computed by statistical analysis of small image windows from large databases [10], [13], [19].

[1]Several terminologies are used in the literature, thought referring to the same task: labeling (*e.g.* [8]), simultaneous segmentation and recognition (*e.g.* [22]), classification (*e.g.* [19]).

A large literature in recent years has been dedicated to modeling the image via a bag of words/features/textons representation [19, 24, 18, 23]. The orderless (independence and exchangeability) assumption of the bag-of-words yields in a simple and easy-to-manipulate representation, *i.e.* vectors or histograms. However, because it overlooks the spatial location of textons in the image, it entails a loss of valuable and discriminative information associated to *pattern and shape features*. To tackle this problem, [12] proposes to take into account the spatial layout of patches through a pyramidal approach. In [16], keypoints extracted from the image are given some importance according to their relative position within the object. The shape-context descriptor [2] is designed to capture the relative position of other shape points around a reference point. In [1], the authors use generalised correlograms to extract information about the neighbourhood of a keypoint.

A classic representation theorem due to De Finetti [5] asserts that any collection of exchangeable random variables can be modeled as a mixture distribution. In this respect, Sivic exploited earlier developments in text classification, document analysis and machine learning (see [17] for example), and proposes a pLSA model (*e.g.* [9]) for image categories classification [22]. Weber introduces a generative mixture model to represent the variability of shape and appearance of objects [25]. A thorough hierarchical approach is developed in [24] to model objects, parts and the scene context.

In this paper, an image is represented as a number of individual *patches*. The size of the patches is taken large enough to enable the computation of robust statistics of textons within each patch, and small enough to assign each patch to a unique *visual class*. Our contribution is threefold. First, inspired from [1], we introduce a new pattern-texton descriptor that takes into account the spatial pattern/layout of points of interest extracted by an Harris detector. We show that classification results from our descriptor outperform results obtained from an appearance-based descriptor alone. Second, we propose a generative model of the patch under the form of mixture components of independent textons. In this model, each class corresponds to a latent variable, which in turn is associated to a set of parameters that define the distribution of textons. Lastly, we show that the optimal number of classes that describe a database can be estimated automatically and exactly by selecting the optimal complexity of the model. This modeling gives very efficient results on a database generated from very high resolution optical images.

The remainder of the paper is organised as follows. In Section 2 we introduce a probabilistic modeling for weakly-supervised classification; the model definition, the parameters learning and the classification task are detailed. Experimental results on remote sensing images are illustrated and evaluated in Section 3. We conclude in Section 4.

2 Probabilistic framework

This section presents a probabilistic framework for classifying highly textured images. We define a probabilistic generative model of the data; three assumptions about the generative process are made: i) the data are described by appearance and shape/pattern characteristics; pattern and appearance features are independently sampled and exchangeable; ii) the probability distribution that generates the data takes the form of a mixture model parametrised by the set $\{\Theta, \pi\}$; iii) there is a one to one correspondence between mixture components and classes (*i.e.* a component is associated to one class only). Each component of the mixture is associated to a latent variable L.

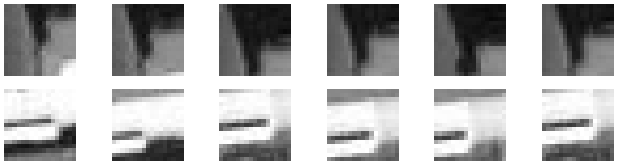

Figure 1: Example of some visual words. Each line corresponds to several different elements of the same visual cluster.

Before giving the details of the generative model (subsection 2.2), we describe the features that are used to feed this model (subsection 2.1). The parameters of the distribution are estimated during the training task, from a weakly labelled database (subsection 2.3). We show how to compute optimally the number of classes that describe the data set (subsection 2.4). Equipped with the estimated parameters, the classification is performed by maximisation of the posterior probability (subsection 2.5).

2.1 Descriptors

Appearance feature Keypoints are extracted using the Harris detector [7]. Amongst the main advantages of this detector are its invariance to rotation and translation. We sample a fixed number of keypoints in a given patch by keeping the N highest responses of the Harris detector, where N is large enough to make possible statistical modeling. As patches of fixed size are considered, we do not make use here of scale adaptation. Following several recent approaches [22, 19], we use SIFT descriptor as appearance feature [14]. SIFT vectors, of dimension 128, are computed on a square of size 16×16 pixels and quantised in n_t cluster using a k-means algorithm. The quantised descriptors, namely *appearance-textons*, constitute a discrete vocabulary $\{t_1, ..., t_{n_t}\}$.

Pattern/shape feature The general idea from the pattern descriptor we propose here is to characterise the spatial distribution of keypoints within a given region by computing the Fourier transform of the function of their occurrences for a given radius. We explain it here as a transformation of a generalised correlogram taken at a keypoint [1].

Thus, let $\{x_1, ..., x_N\}$ be the set of vectors of location of keypoints extracted in the patch, and $\{p_1, ..., p_N\}$ the textons extracted at these points. At each keypoint x_i, a generalised correlogram h_i is extracted as follows [1]:

Let the spatial relation $(x_i - x_j)$ of any keypoint x_j with reference to x_i be written in polar coordinates (α_{ij}, r_{ij}), where $\alpha_{ij} = \widehat{(x_i - x_j)}$ and $r_{ij} = \|x_i - x_j\|$. The angle and radius are quantised in n_α and n_r bins respectively. Let note A_u be the u-th bin of angles, $u \in \{1, 2, ..., n_\alpha\}$, and let R_v be the v-th bin of radius, $v \in \{1, ..., n_r\}$. The generalised correlogram h_i, of dimension $n_h = n_\alpha \times n_r \times n_t$ and linked to p_i, is defined by:

$$h_i(t,u,v) = \frac{1}{N} \left| \{ p_j \in P : p_j = t, \widehat{(x_i - x_j)} \in A_u, \|x_i - x_j\| \in R_v \} \right| .$$

Taken as such, the correlograms have very high dimensionality, are usually very sparse, and are not rotation-invariant. To tackle these problems, we propose to apply to the correlogram the following transformation: Given a bin of radius v and a keypoint x_i, we

consider the unidimensional function $f_{iv}(u) = \sum_{t=1}^{n_t} h_i(u,v,t)$; this function simply counts the occurrences of textons in every bin of angle for a given bin of radius, independently of the texton bin. We compute the Fourier transform of f_{iv} and retain the module of the n_f first coefficients. It results in a descriptor which is invariant to rotation transformation, and of size $n_f \times n_v < n_h$. Finally, we build, from a large database, a *pattern-textons codebook* of size n_s. Then, to each keypoint of the image is assigned an entry in the pattern codebook.

2.2 Probabilistic modeling

We assume that each patch of index i is linked to the realisation of a discrete *latent variable* L_i, which is taken independently for each patch. This latent variable takes its values in a vocabulary $\{V^1, ..., V^K\}$ of size K, with probability $\pi = \{\pi_1, ..., \pi_K\}$ respectively; these values correspond to the different *classes* which will be assigned to the patches.

We consider that appearance-textons T_i and pattern-textons S_i in the patch i depend only on the latent variable L_i. We make also the strong assumption that S_i and T_i are independent conditionally to L_i, *i.e.* : $P(S_i, T_i|L_i) = P(S_i|L_i)P(T_i|L_i)$. Moreover, we suppose that the textons within the set of textons S_i are independent conditionally to L_i; the same assumption is made for the set T_i.

To simplify the notations, we concatenate the representations of S_i and T_i in a single histogram vector O_i. If n_T and n_S are respectively the number of quantisation values of the SIFT vectors and the pattern feature descriptors, the total histogram is of size $n = n_s + n_t$. The vocabulary of this set of textons is given by $\{o_1, ..., o_n\}$.

Let o be a texton inside this patch, we define $\theta_{jk} = P(o = o_j|L_i = V^k)$ the probability of the realisation of texton entry o_j, conditionally to V_k. The distributions $P(O_i|L_i)$ are then entirely described by the set of parameters $\Theta = \{\theta_{jk}\}_{1 \leq j \leq n, 1 \leq k \leq K}$.

The observation $O_i = \{N_{i1}, N_{i2}, ..., N_{in}\}$ stands for the histogram of the textons inside the patch i, where N_{ij} is the number of occurrences of entry o_j in this patch. Using the assumption of independence of textons conditionally to the latent variable, the probability to generate the textons in patch i writes:

$$P(O_i|\Theta, \pi) = \sum_{k=1}^{K} P(L_i = V_k|\Theta, \pi) P(O_i|L_i = V_k) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{n} \theta_{jk}^{N_{ij}} . \quad (1)$$

Thus, each visual class of a database is associated to a latent variable, which in turn is associated to one component of the mixture model. This modelisation is a mixture of unigrams model ([17]) and, unlike pLSA, each patch is supposed to be associated to a single value of the latent variable.

2.3 Learning the probability distribution

The learning task consists of estimating the parameters that defines the mixture model (1), using a training set of unlabelled data containing K visual classes. These parameters are estimated by maximising the likelihood of the observations $P(O|\Theta, \pi)$.

The Expectation-Maximization algorithm [6] suits perfectly this optimisation problem: it gives an efficient way to compute step by step a (local) maximum of the likelihood, in case of incomplete data. Our data is considered incomplete because it comes without a class label.

We introduce a hidden variable z —which can be associated to the indicator function of the latent variable. The following two steps are computed iteratively and until convergence:

E-step: computation of the expectation $\gamma_k(O_i) = E(z_{ik}|O_i,\Theta,\pi)$, for all k and i, using the Bayes inversion rule (t stands for the iteration index):

$$\gamma_k^{(t+1)}(O_i) = \frac{\pi_k^{(t)} \prod_{j=1}^{n} (\theta_{jk}^{(t)})^{N_{ij}}}{\sum_{k=1}^{K} \pi_k^{(t)} \prod_{j=1}^{n} (\theta_{jk}^{(t)})^{N_{ij}}}$$

M-step: maximisation of $E_{Z|O,\Theta,\pi}(\log P(O,z|\Theta,\pi))$, using the Lagrange multipliers method. The following update formula are thus obtained:

$$\left\{ \begin{array}{l} \theta_{jk}^{(t+1)} = \frac{\sum_{i=1}^{N} \gamma_k^{(t)+1}(O_i) n_i}{\sum_{i=1}^{n} \gamma_k^{(t+1)}(O_i)\, N_t} \\ \pi_k^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} \gamma_k^{(t+1)}(O_i) \end{array} \right. \tag{2}$$

with N the total number of patches in the learning database, and N_t the total number of textons in a patch.

The estimated parameters at the end of the iterations are noted $\{\tilde{\Theta}', \tilde{\pi}'\}$.

Initialization of the parameters The resulting estimated components depend upon the initialization of the parameters $\{\Theta^{(t=0)}, \pi^{(t=0)}\}$ at the first E-step. In order to ensure a *robust convergence* (in the sense of repetitiveness; it does not ensure a *global* convergence), we first apply a k-means clustering algorithm to the normalised $O_{i\in\{1,\ldots N\}}$ histograms (here taken as vectors); it results in a preliminary estimation of the components $\tilde{\Theta}$. The EM is then initialized by assigning $\Theta^{(t=0)} = \tilde{\Theta}$. Weighting parameters π are initialised from a uniform distribution. This process is repeated several times. Finally, the optimal set of parameters that is used during classification is defined by: $\{\hat{\Theta}, \hat{\pi}\} = \arg \max_{\{\tilde{\Theta}', \tilde{\pi}'\}} \prod_i P(L_i|O_i, \tilde{\Theta}', \tilde{\pi}')$, where O_i is computed from the training database.

2.4 Model selection

Fixing arbitrarily and *a priori* the number of classes K that describe a given database is not quite satisfying. We show in this subsection how to exploit the Minimum Description Length principle to compute automatically the *optimal number of classes* of a training database.

Introduced in 1989 [20], the *Minimum Stochastic Complexity criterion* assumes that the best model describing a database is the one coding it with a minimal number of bits. Given a set of models $\{M_K\}_{K=1}^{K_{max}}$ and the learning database O, the model M_K is associated to the description length $D(O,M_K)$. This length is separated into two parts, the length of the code which is necessary to code a model, and the length of the code necessary to code the data using the model: $D(O,M_K) = D(O|M_K) + D(M_K)$.

In [21], Shannon establishes a relation between the code length and the probability of the signal: $D(O|M_K) = -\log P(O|M_K)$. Incorporating Equation (1), this expression becomes:

$$D(O|M_k) = -\sum_{i=1}^{N} \log(\sum_{k=1}^{K} \hat{\pi}_k \prod_{k=1}^{n} \hat{\theta}_{jk}^{N_{ik}}) \,. \tag{3}$$

The model length $D(M_K)$ should, in theory, be infinite, because M_K consists of a set of real parameters. However, since the parameters of our model are estimated from the number of occurrences of textons in the patches (see Equ. 2), we can use a formula given by Rissanen in [20] and estimate $D(M_K)$ by:

$$D(M_K) = \frac{K}{2}[(n+1)log(N) + nlog(N_t)] \,, \tag{4}$$

where N is the total number of patches in the learning database, N_t is the total number of textons in all patches, and n is the size of the descriptor.

In order to find the optimal model, we use the following heuristic: for all K varying from 1 to a fixed value K_{max}, the parameters of the models M_K are estimated and the stochastic complexity $D(O, M_K)$ is computed. The model M_K that minimises the stochastic complexity is retained and determines the optimal number of classes K^{opt}.

2.5 Classification

Given estimates of the parameters $\{\hat{\Theta}, \hat{\pi}\}$, the task is to classify a patch of the test data into a single class. The selected class will be the one that is associated to the latent variable which has the highest posterior probability $P(L_i|O_i, \hat{\Theta}, \hat{\pi})$. For this purpose, we apply a Bayes decomposition rule:

$$P(L_i|O_i, \hat{\Theta}, \hat{\pi}) = \frac{P(O_i|L_i, \hat{\Theta}, \hat{\pi})P(L_i|\hat{\Theta}, \hat{\pi})}{P(O_i|\hat{\Theta}, \hat{\pi})} \,. \tag{5}$$

Note that the denominator of Equ. (5) does not depend on L_i. The selected class on patch i is then given by:

$$\arg \max_{V^k \in \{1,\ldots,K\}} P(L_i = V^k|O_i, \hat{\Theta}, \hat{\pi}) = \arg \max_{V^k \in \{1,\ldots,K\}} P(O_i|L_i = V^k, \hat{\Theta}, \hat{\pi})P(L_i = V^k|\hat{\Theta}, \hat{\pi}) \,, \tag{6}$$

with $P(O_i|L_i = V_k, \hat{\Theta}, \hat{\pi})P(L_i = V_k|\hat{\Theta}, \hat{\pi}) = \hat{\pi}_k \prod_{j=1}^{n} \hat{\theta}_{jk}^{N_{ij}}$.

It is possible to introduce a reject class by setting a threshold value *thresh* on the posterior. It means that the patch will be rejected if no class corresponds to a sufficient likelihood. Thus, a patch is not classified if:

$$\max_{k \in \{1,\ldots,K\}} P(L_i = V^k|O_i) < thresh \,. \tag{7}$$

3 Experimental results on satellite images

In this section, we describe the composition of our data sets, the experimental setup, and the results obtained for the tasks of classification and model selection. We compare classification results obtained with and without the pattern descriptor and show evidence of the improvement of the classification in the former case.

3.1 Data sets and features

Our data set is generated from an optical very high resolution Quickbird panchromatic image. The original Quickbird image being very large ($\approx 10,000^2$ pixels) we extracted

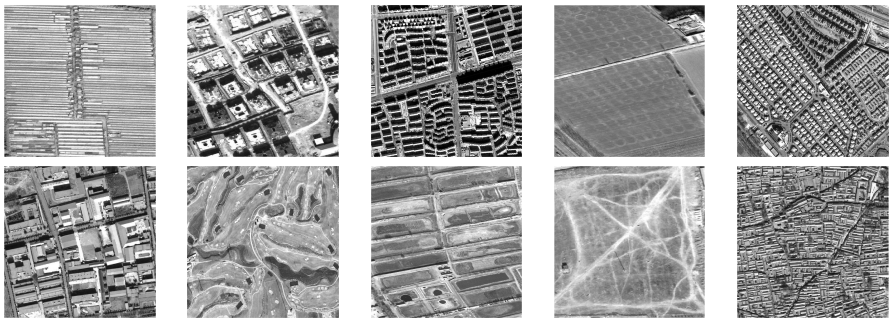

Figure 2: Instances of the visual classes. From left to right: Greenhouses, Working place, Big-building area, Fields, Housing area, Small industries area, Golf field area, Fishing area, Wasteland, Hutong.

5535 patches, of size 250×250 pixels each, which we divided into two distinct sets: the learning set (containing 3823 patches), and the test set (containing 1712 patches). The whole data set comprises 10 visual classes illustrated in Figure 2.

Appearance and pattern textons (in total around 1 million vectors) are computed from images which are included neither in the training data nor in the testing data set. We generate a SIFT feature codebook of size $n_t = 180$. For the pattern descriptor, we set $(R, n_r, n_\alpha, n_f) = (30, 3, 8, 4)$. Each patch of the data base is then represented by a histogram of size $n = 225$.

3.2 Classification

The classification task was performed and quantitatively evaluated on a data set containing 8 visual classes (see Tab. 1) —taken out of the 10 of the original database. During the learning stage, EM algorithm was applied with 30 different initialisations, assuming that the number of classes $K = 8$ is known. During the test stage, patches of the test set are classified from Equation (6).

Classification results are illustrated in Tabular 1. To analyse the effect of pattern-textons, we compared different descriptors: SIFT textons, SIFT + pattern textons, and Haralick features. Haralick features —texture features computed from a co-ocurrence matrix— have proved to be powerful for classification in images of 2.5m resolution [3].

It appears clearly that the performance of our classifier improves significantly when we incorporate pattern-textons, especially on classes containing a lot of geometrical information (see Tab. 1): the correct classification rate can increase up to $\sim 14\%$ when adding pattern-textons to the SIFT textons. We observe also that Haralick features are always less performing than our pattern-descriptor —the average correct classification is about 77% with Haralick, against 95% for our descriptor–, except for one class. It is worth noticing that we did not try to tune the parameter set of our pattern-descriptor $(R, n_r, n_\alpha, n_f, m)$: we could verify experimentally that certain parameters better fit certain geometric patterns, however it is out of the scope of this paper.

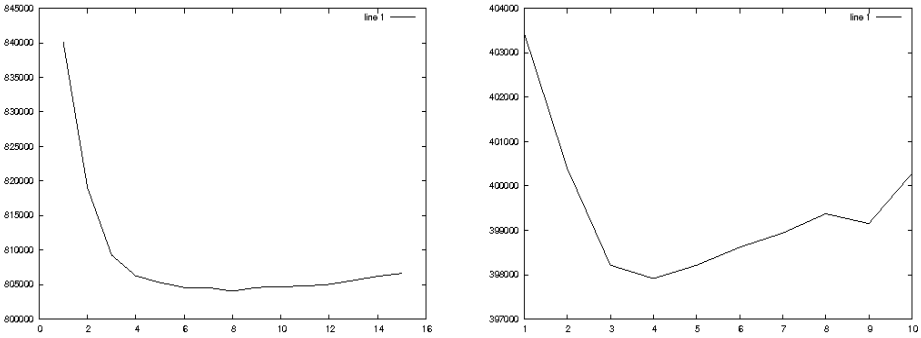

Figure 3: (a): Stochastic complexity with 8 classes (left) and with 4 classes (right). The minimum complexity corresponds exactly to the expected number of classes.

3.3 Model selection

We applied our model selection algorithm on two different training data sets containing respectively K=4 classes and K=8 classes. The model that describes best our data is given at the minimum stochastic complexity (Equ. (3) and (4)). Note that applying model selection is equivalent to learning the (optimal) number of classes in a database.

Results are illustrated in Figure 3. As expected, the minimal description length corresponds to $K^{opt} = 4$ for the first data set, and to $K^{opt} = 8$ for the second: in both cases the number of classes is correctly estimated.

This result indicates that our probabilistic model, defined as mixture distributions of independent textons, correctly reflects the data.

3.4 Classification using a reject class

We tested our classification procedure on a large size image. Parameters $\hat{\Theta}$ and $\hat{\pi}$ were estimated on a small training set (around 500 patches) of 5 visual classes (classes (a), (d), (e), (f), (j) from Fig. 2). We assume that the test image to be classified might contain some regions which do not belong to any of the predefined classes; we therefore introduce a reject threshold on the posterior probability (see Equ. (7)), whose value can be deduced from the observation of the distribution of the posterior on training data. A regular grid divides the entire test image in patches of size 250×250 pixels. Each patch is classified independently using Equation (6).

Qualitative results are illustrated in Figure 4. They are consistent with expected results. We observe —-visually and by detailed inspection of the estimated posterior for each of the classes— that the rejected patches correspond either to patches containing a mixture of predefined classes, or to patches that should belong to a class which has not been learned. We notice the spatial homogeneity of the classification.

Visual classes	CCR *without* pattern descriptor	CCR *with* pattern descriptor	CCR with *Haralick* features
(a) Green houses	93.94	93.94	95.68
(b) Work place	84.62	98.56	47.68
(c) Big building	87.30	89.68	73.29
(e) Housing area	98.82	99.29	93.37
(f) Small industr.	84.21	94.77	52.97
(g) Golf field	96.40	96,40	94.53
(h) Fishing area	92.94	92.30	86.51
(j) Hutong	98.15	98.41	98.22
Av. CCR	92.36	95.63	77.32

Table 1: Performance evaluation and comparison from 8 visual classes (see Fig. 2). CCR stands for the *Correct Classification Rate*. Our pattern-descriptor improves significantly the correct classification rate for classes (b) and (f), and performs always better than Haralick features, except for class (a).

4 Conclusion

We presented in this paper a probabilistic modeling to classify remote sensing images. We introduced a new rotation invariant pattern descriptor. The distribution of textons in a patch is modeled by a mixture distribution, based on the assumption of textons' independence and exchangeability. We showed that the optimal number of classes to describe a database can be computed exactly by minimising the stochastic complexity of the model. Very satisfying results are obtained on a database of Quickbird images. Future work will be dedicated to developing a multi-scale version of our model and integrating it in a CRF framework.

References

[1] J. Amores, N. Sebe, and P. Radeva. Object-class recognition and retrieval by generalized correlograms. *PAMI*, 29, 2007.

[2] S. Belongie, J. Malik, and J. Puzicha. Shape context: a new descriptor for shape matching and object recognition. In *Proc. NIPS*, 2000.

[3] M. Campedel, B. Luo, H. Maitre, E. Moulines, M. Roux, and I. Kyrgyzov. Indexation des images satellitaires. Technical report, ENST, 2004.

[4] M. Datcu, H. Daschiel, A. Pelizzari, and al. Information mining in remote sensing image archives: system concepts. *TGRS*, 41(12), 2003.

[5] B. de Finetti. *Theory of probability*. John Wiley and Sons Ltd, Chichester, 1975.

[6] T. Dempster. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39, 1977.

[7] C. Harris and M. Stephens. A combined edge and corner detector. In *Proc. 4th Alvey Vision Conference*, 1988.

[8] X. He, R.S. Zemel, and M.A. Carreira-Perpinan. Multi-scale conditional random fields for image labeling. In *Proc. CVPR*, 2004.

Figure 4: Classification results on a Quickbird image of size 2250×3500 pixels. Each color corresponds to a class; white colour indicates the rejected patches.

[9] T. Hofmann. Probabilistic latent semantic indexing. In *Proc. SIGIR*, 1999.

[10] J. Huang. *Statistics of natural images and models*. PhD thesis, Division of applied Mathematics, Brown University, 2000.

[11] B. Julesz. Textons, the elements of texture perception, and their interactions. *Nature*, 290, 1981.

[12] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. CVPR*, 2006.

[13] A. Lee, K. Pedersen, and D. Mumford. The non-linear statistics of high-contrast patches in natural images. *IJCV*, 54, 2003.

[14] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60, 2004.

[15] D. Marr. *Vision*. Freeman Publisher, 1983.

[16] M. Marszalek and C. Schmid. Spatial weighting for bag-of-features. In *Proc. CVPR*, 2006.

[17] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled document using EM. *Machine Learning*, 39, 2000.

[18] E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In *Proc. ECCV*, 2006.

[19] P. Quelhas, F. Monay, J.M. Odobez, D. Gatica-Perez, and T. Tuytelaars. A thousand words in a scene. *PAMI*, 29, 2007.

[20] J. Rissanen. Modeling by shortest data description. *Automatica*, 14, 1978.

[21] C. Shannon. A mathematical theory of communication. *Bell Syst Technology*, 27, 1948.

[22] J. Sivic, B.C. Russell, A.A. Efros, A. Zisserman, and W.T. Freeman. Discovering objects and their location in images. In *Proc. ICCV*, 2005.

[23] J. Sivic and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proc. CVPR*, 2007.

[24] E.B. Sudderth, A. Torralba, W.T. Freeman, and A.S. Willsky. Describing visual scenes using transformed objects and parts. *IJCV*, 77, 2007.

[25] M. Weber, M. Welling, and P. Perona. Towards automatic discovery of object categories. In *Proc. CVPR*, 2000.