IDIAP RESEARCH REPORT

INTEGRATING CO-OCCURRENCE AND SPATIAL CONTEXTS ON PATCH-BASED SCENE SEGMENTATION

Florent Monay [a]   Pedro Quelhas [a]
Jean-Marc Odobez [a]   Daniel Gatica-Perez [a]

IDIAP–RR 05-30

JUNE 2005

REVISED IN SEPTEMBER 05

PUBLISHED IN
Beyond Patches Workshop, in conjuction with CVPR 2006

[a] IDIAP Research Institute, 1920 Martigny,Switzerland

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 05-30

Integrating co-occurrence and spatial contexts on patch-based scene segmentation

Florent Monay Pedro Quelhas Jean-Marc Odobez Daniel Gatica-Perez

June 2005

revised in September 05

published in
Beyond Patches Workshop, in conjuction with CVPR 2006

Abstract. We present a novel approach for contextual segmentation of complex visual scenes, based on the use of bags of local invariant features (visterms) and probabilistic aspect models. Our approach uses context in two ways: (1) by using the fact that specific learned aspects correlate with the semantic classes, which resolves some cases of visual polysemy, and (2) by formalizing the notion that scene context is image-specific -what an individual visterm represents depends on what the rest of the visterms in the same bag represent too-. We demonstrate the validity of our approach on a man-made vs. natural visterm classification problem. Experiments on an image collection of complex scenes show that the approach improves region discrimination, producing satisfactory results, and outperforming a non-contextual method. Furthermore, through the later use of a Markov Random Field model, we also show that co-occurrence and spatial contextual information can be conveniently integrated for improved visterm classification.

Figure 1: Scene segmentation using local invariant regions (yellow). Regions are classified either as man-made (blue) or nature (not shown), and superimposed on a manual segmentation (white).

1 Introduction

Associating semantic class labels to image regions is a fundamental task in computer vision, useful in itself for image, video indexing and retrieval, and as an intermediate step for higher-level scene analysis [6, 8, 18, 19]. While most segmentation approaches segment image pixels or blocks based on their luminance, color or texture, in this work we consider local image regions characterized by viewpoint invariant descriptors [10]. This region representation, robust with respect to partial occlusion, clutter, and changes in viewpoint and illumination, has shown its applicability in a number of vision tasks [2, 3, 8, 14–16, 20]. Although local invariant regions do not provide a full segmentation of an image, they often occupy a considerable part of the scene and thus can define a "sparse" segmentation (Fig. 1).

In general, the constituent parts of a scene do not exist in isolation, and the visual context -the spatial dependencies between scene parts- can be used to improve region classification [6,7,9,13]. Two regions, indistinguishable from each other when analyzed independently, might be discriminated as belonging to the correct class with the help of context knowledge. Broadly speaking, there exists a continuum of contextual models for image segmentation. On one end, one would find explicit models like Markov Random Fields, where spatial constraints are defined via local statistical dependencies between class region labels [4,9], and between observations and labels [6]. The other end would correspond to context-free models, where regions are classified assuming statistical independence between the region labels, and using only local observations [2].

Lying between these two extremes, a type of representation of increasing use is the bag-of-visterms (BOV), i.e., a histogram of discretized regional descriptors. On one hand, unlike explicit contextual models, spatial neighboring relations in this representation are discarded, and any ordering between the descriptors disappears. On the other hand, unlike point-wise models, although the descriptors are still local, the scene is represented collectively. This can explain why, despite the loss of "strong" spatial contextual information, BOVs have been successfully used in a number of problems, including object matching [16] and categorization [15,20], scene classification [3,14] and retrieval [18].

As a collection of discrete data, the BOV representation is suitable for probabilistic models where a different form of context is implicitly captured through visterm co-occurrence. These models, originally designed for text collections (documents composed of terms), use a discrete hidden *aspect* variable to model the co-occurrence of terms within and across documents. Examples include Probabilistic Latent Semantic Analysis (PLSA) [5] and Latent Dirichlet Allocation (LDA) [1]. We have recently shown that the integration of PLSA and BOVs defined on invariant local descriptors can be successfully used for global scene classification [14]. Given an unlabeled image set, PLSA captures aspects that represent the class structure of the collection, and provides a low-dimensional representation useful for classification. Similar conclusions with an LDA related model were reached in [3].

The main issue with classifying regions using visterms is that they are not class-specific. As shown in Fig. 2, the same visterms commonly appear both in man-made and nature views. This situation, although expected since visterm construction usually does not make use of class label information,

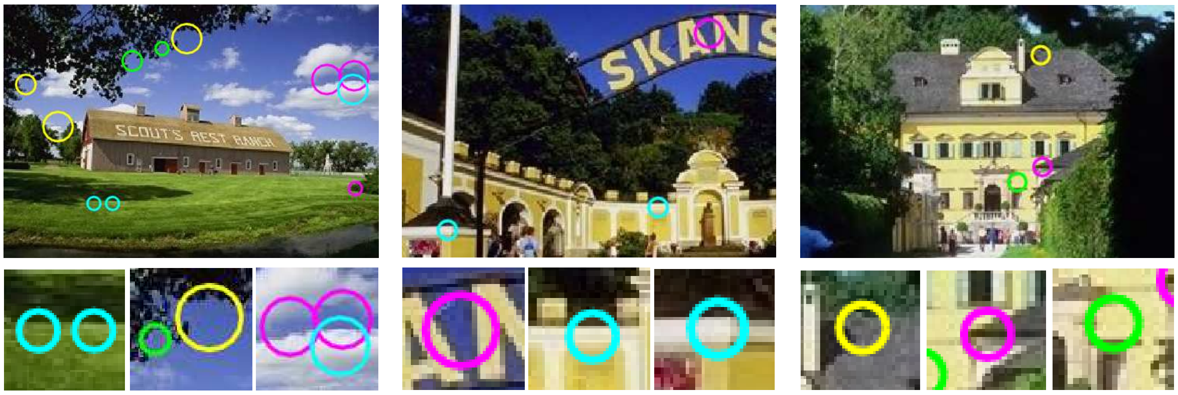

Figure 2: Regions (represented by visterms) can have different class labels depending of the images where they are found. Left: various regions (4 different colors, same color means same visterm) that occur on *natural* parts of an image. Center Right: the same visterms occur in man-made structures. All these regions were correctly classified by our approach, switching the class label for the same visterms depending on the context.

constitutes a problematic form of visual polysemy. In this paper, we show that aspect models can also be used for region classification. We propose probabilistic models that exploit two ways of using context. In the first place, we use the fact that specific learned aspects correlate with the semantic classes, which implicitly helps in cases of polysemy [5]. In the second place, scene context is image-specific: the "meaning" of a particular visterm depends on what the "meaning" of the other visterms in the same bag is. We show that this relation can be formally expressed in the probabilistic model, so that even though visterms occur in both classes, the information about the other visterms in the same bag can be used to improve discrimination (Fig. 2).

In this paper we propose two probabilistic aspect models that learn this co-occurrence information for visterm classification. We present results on a man-made vs. natural region classification task, and show that the contextual information learned from co-occurrence improves the performance compared to a non-contextual approach. In our view, the proposed approach constitutes an interesting way to model visual context that could be applicable to other problems in computer vision. Furthermore, we show, through the use of a Markov Random Field model, that standard spatial context can be integrated, resulting in an improvement of the final segmentation.

The paper is organized as follows. Section 2 reviews the closest related work. Section 3 introduces our model for contextual scene segmentation. Section 4 reports our results, including the MRF modeling. Section 5 concludes the paper.

2 Related work

There is an abundant literature on image segmentation. The perspective on image segmentation that we consider in this paper differs from the traditional notion of homogeneous region partion of the image. We perform segmentation of the image based on class labels defined over the whole database, and we base our segementation on the classification of local patches that do not cover the whole image. In this section we briefly consider some of the related work that is most relevant to our approach.

In [2], invariant local descriptors are used for an object segmentation task. All region descriptors that belong to the object class in the training set are modeled with a Gaussian Mixture Model (GMM), and a second GMM is trained on non-object regions. In this non-contextual approach, new regions are independently classified depending on their relative likelihood with respect to the object and non-object models. A similar approach introducing spatial contextual information through neighborhood statistics of the GMM components collected on training images is proposed in [8], where the learned prior statistics are used for relaxation of the original region classification.

In image segmentation, quantized local descriptors -referred to as *textons*- have also been used to build local BOV representations of windowed image regions [11]. The similarity between these regions is then defined based on this histogram representation, and segmentation is conducted for each individual image using spectral clustering.

Exploring spatial dependencies, Kumar and Herbert apply a random field model to segment image areas that represent man-made scene structures [6]. Their approach is based on the extraction of features from a grid of block that fully cover the image.

Using a similar grid layout, Vogel and Schiele recently presented a two-stage framework to perform scene retrieval [18] and scene classification [19]. This work involves an intermediate image block classification step, that can be seen as scene segmentation.

Probabilistic aspect models have been recently proposed to capture visterm co-occurrence information with the use of a hidden variable (latent aspect). The work in [3] proposed a hierarchical Bayesian model that extended LDA for global categorization of natural scenes. This work showed that important visterms for a class in an image can be found. However, the problem of region classification for scene segmentation was not addressed. The combination of local descriptors and PLSA for image segmentation has been illustrated in [15] and [14]. However these works have two limitations. First, visterms were classified into aspects, not classes, unless we assume as in [15] that there is a direct correspondence between aspects and semantic classes. This seems however a quite unrealistic assumption in practice, since it implicitely assumes a one-to-one correspondence between aspects and class labels. In [14], an ad-hoc procedure was used to relate aspects and classes, in which a class would be represented by the aspects resulting in the best average precision on an image retrieval task. Secondly, evaluation was limited, e.g. [15] does not conduct any objective evaluation of the segmentation performance.

Unlike these previous approaches, we propose a formal way to integrate the latent aspect modeling in the class information, and conduct a proper performance evaluation, validating our work with a comparison to a state-of-the-art baseline method. In addition, we explore the integration of the more traditional spatial MRFs into our system and compare the obtained segmentations.

3 Patch-based scene segmentation

Our segmentation task can be formulated as the automatic extraction of patches (referred to as visterms in the paper) from the image followed by the classification of each visterm into a class c, where c stands either for *m*an-made structures or *natural* regions. In the next subsections, we first focus on the classification models, and then summarize the visterm extraction process.

Assume a discrete set of image patches (visterms), corresponding to the quantization of local descriptors (see Section 3.3). We rely on likelihood ratio computation to classify each visterm v of a given image d into a class c. The ratio is defined by

$$LR(v) = \frac{P(v|c = \text{man-made})}{P(v|c = \text{natural})}, \tag{1}$$

where the probabilities will be estimated using different models of the data, as described in the next subsections.

3.1 Empirical distribution

Given a set of training data, the term in Eq.1 can simply be estimated using the empirical distribution of visterms, as done in [2]. More precisely, given a set of manually segmented images $\mathcal{D}$into man-made and natural regions (e.g. Fig.1 (c)), $P(v|c)$ is estimated as the number of times the visterm v appears in regions of class c, divided by the number of occurrences of v in the training set.

3.2 Aspect modeling

Empirical estimation of probabilities is simple but may suffer from several drawbacks. A first one is that a significantly large amount of training data might be necessary to avoid noisy estimates, especially when using large vocabulary sizes. A second one is that such estimation only reflects the individual visterm occurrences, and does not account for any kind of relationship between them. We propose to exploit aspect models [1, 5] that capture visterm co-occurrences to classify visterms. These models, through the identification of latent aspects, enable the classification of the visterms of one image based on the occurrence of other visterms in the same image. The histogram of visterms in image d, referred to as the bag-of-visterms (BOV), contains this information. Even if the BOV representation discards all spatial neighboring relations, we expect the co-occurrence context (i.e. the other visterms) to help for the classification of individual visterms. To this end, we propose two models.

3.2.1 Aspect model 1

The first model associates a hidden variable $z \in \mathcal{Z} = \{z_1, \dots z_{N_A}\}$ with each observation according to the graphical model of Fig. 3, leading to the joint probability defined by

$$P(c, d, z, v) = P(c)P(d|c)P(z|d)P(v|z). \tag{2}$$

This model introduces several conditional independence assumptions. The first one, traditionally encountered in aspects models, is that the occurrence of a visterm v is independent of the image d it belongs to, given an aspect z. The second assumption is that the occurrence of aspects is independent of the class the document belongs to. The parameters of this model are learned using the maximum likelihood (ML) principle [5]. The optimization is conducted using the Expectation-Maximization (EM) algorithm, allowing us to learn the aspect distributions $P(v|z)$ and the mixture parameters $P(z|d)$.

Notice that, given our model, the EM equations do not depend on the class label. Besides, the estimation of the class-conditional probabilities $P(d|c)$ does not require the use of the EM algorithm. We will exploit these points to train the aspect models on a large dataset (denoted $\mathcal{D}$) where only a small part of it has been manually labeled according to the class (we denote this subset by $\mathcal{D}_{lab}$). This allows for the estimation of a precise aspect model, while alleviating the need for tedious manual labeling. Regarding the class-conditional probabilities, as the labeled set will be composed of man-made-only or natural-only images, we simply estimate them according to:

$$P(d|c) = \begin{cases} 1/N_c & \text{if } d \text{ belongs to class } c \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where N_c denotes the number of images belonging to class c in the labeled set $\mathcal{D}_{lab}$. Given this model, the likelihood we are looking for can be expressed as

$$P(v|c) = \sum_{l=1}^{N_A} P(v, z_l|c) = \sum_{l=1}^{N_A} P(v|z_l)P(z_l|c), \tag{4}$$

where the conditional probabilities $P(z_l|c)$ can in turn be estimated through marginalization over labeled documents,

$$P(z_l|c) = \sum_{d \in \mathcal{D}_{lab}} P(z_l, d|c) = \sum_{d \in \mathcal{D}_{lab}} P(z_l|d)P(d|c). \tag{5}$$

These equations allow us to estimate the likelihood ratio as defined by Eq.1. Note that this model extends PLSA by introducing the class variable [5].

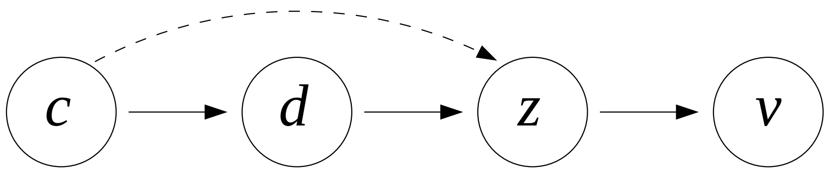

Figure 3: Aspect model 1 and aspect model 2 (dashed line).

3.2.2 Aspect model 2

From Eq. 4, we see that, despite the fact that the above model captures co-occurrence of the visterms in the distributions $P(v|z)$, the context provided by the specific image d has no direct impact on the likelihood. To explicitly introduce this context knowledge, we propose to evaluate the likelihood ratio of visterms conditioned on the observed image d,

$$LR(v, d) = \frac{P(v|d, c = \text{man-made})}{P(v|d, c = \text{natural})}. \tag{6}$$

The evaluation of $P(v|d, c)$ can be obtained by marginalizing over the aspects,

$$P(v|d, c) = \sum_{l=1}^{N_A} P(v, z_l|d, c) = \sum_{l=1}^{N_A} P(v|z_l)P(z_l|d, c), \tag{7}$$

where we have exploited the conditional independence of visterm occurrence given the aspect variable. Under model 1 assumptions, $P(z_l|d, c)$ reduces to $P(z_l|d)$, which clearly shows the limitation of this model to introduce both context and class information. To overcome this, we assume that the aspects depend on the class label as well (cf dashed link in Fig. 3). The parameters of this model are the aspect multinomial $P(v|z)$ and the mixture multinomial $P(z|d, c)$, which could be estimated from labeled data by EM as before. However, as our model is not fully generative [1], only $P(v|z)$ can be kept fixed, and we would have to estimate $P(z|d_{new}, c)$ for each new image d_{new}. Since the class is obviously unknown for new images, this means that in practice all the dependencies between aspects and labels observed in the training data would be lost. To avoid this, we propose to separate the contributions of the aspect likelihood due to the class-aspect dependencies, from the contributions due to the image-aspect dependencies. Thus, we propose to approximate $P(z_l|d, c)$ as

$$P(z_l|d, c) \propto P(z_l|d)P(z_l|c), \tag{8}$$

where $P(z_l|c)$ is still obtained using Eq. 5. The complete expression is given by

$$P(v|d, c) \propto \sum_{l=1}^{N_A} P(v|z_l)P(z_l|c)P(z_l|d). \tag{9}$$

The main difference with Eq.4 is the introduction of the contextual term $P(z_l|d)$, which means that visterms will not only be classified based on their class-likely aspects, but also on the specific occurrence of these aspects in the given image.

Inference on new images

With aspect model 1 (and also with empirical distribution), visterm classification is done once for all at training time, through the visterm co-occurrence analysis on the training images. Thus, for a new image d_{new}, the extracted visterms are directly assigned to their corresponding most likely label. For aspect model 2, however, the likelihood-ratio $LR(v, d_{new})$ (Eq. 6) involves the aspect parameters $P(z|d_{new})$ (Eq. 9). Given our approximation (Eq. 8), these parameters have to be inferred for each new image, in a similar fashion as for PLSA [5]. $P(z_l|d_{new})$ is estimated by maximizing the likelihood of the BOV representation of d_{new}, fixing the learned $P(v|z_l)$ parameters in the Maximization step.

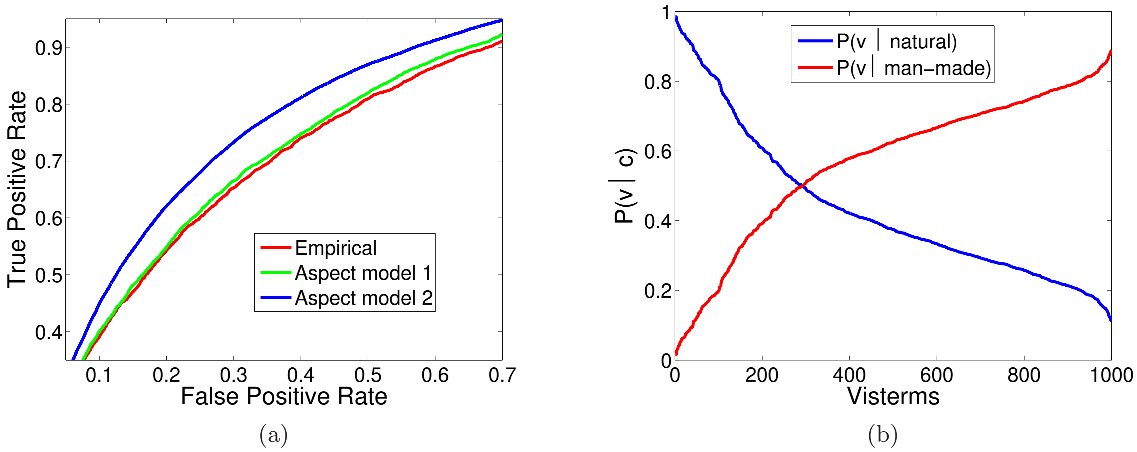

Figure 4: (a) True Positive Rate vs. False Positive Rate for the three methods. (b) $P(v \mid c)$ for man-made and natural structures, estimated on the test set.

3.3 Visterm extraction

Three steps are involved in the construction of the BOV representation: (i) detection of interest points, (ii) computation of local descriptors, (iii) local descriptor quantization. Different point detectors have been proposed to extract regions of interest in images [10, 12, 17]. We use the difference-of-Gaussians (DOG) [10] as point detector, which identifies blob-like regions invariant to translation, scale, rotation, and constant illumination variations. As local descriptors, we use SIFT (Scale Invariant Feature Transform) features [10], which define an orientation-invariant descriptor based on the image grayscale representation. Finally, visterms are obtained by quantizing each local descriptor into an element of a finite vocabulary $\mathcal{V}$, according to a nearest neighbor rule. The vocabulary is obtained by applying the K-means algorithm to a set of local descriptors extracted from the training images, and keeping the means as visterms. We use the Euclidean distance in the clustering and quantization processes. Using the vocabulary $\mathcal{V}$, we attribute to each local descriptor the label of the closest cluster, its corresponding visterm. The final step of the BOV representation is the histograming of the visterms in each image. We obtain the BOV representation from the obtained visterms according to:

$$h(d) = (h_i(d))_{i=1..N_{\mathcal{V}}}, \text{ with } h_i(d) = \mathrm{n}(d, v_i) \tag{10}$$

where $\mathrm{n}(d, v_i)$ denotes the number of occurrences of visterm v_i in image d.

4 Experiments and discussion

We validate our proposed models on the segmentation of scenes into natural vs. man-made structures. This Section first presents our experimental setup. It is followed by a detailed, objective performance evaluation illustrated with segmentation results on a few test images. Finally, we study the integration of a regularization strategy to further improve the segmentation performance.

4.1 Experimental setup

Datasets: Three image subsets from the *Corel Stock Photo Library* were used in the experiments. The first set, $\mathcal{D}$, contains 6600 photos depicting mountains, forests, buildings, and cities. From this set, 6000 have no associated label, while the remaining subset $\mathcal{D}_{lab}$ is composed of 600 images, whose content mainly belonged to one of the two classes, which were hand-labeled with a single class label

	Emp. distr.	Aspect mod. 1	Aspect mod. 2
HTRR	67.5	68.5	72.4

Table 1: Half Total Recognition Rate (in percent).

leading to approximately 300 images of each class. $\mathcal{D}$ was used to construct the vocabulary and learn the aspect models, while $\mathcal{D}_{lab}$ was used to estimate the visterm likelihoods for each class. A third set $\mathcal{D}_{test}$, containing 485 images of man-made structures in natural landscapes, which were hand-segmented with polygonal shapes (Fig. 1), was used to test the methods.

Performance evaluation: The global performance of the algorithm was assessed using the True Positive Rate (TPR, number of positive visterms retrieved over the total number of positive descriptors), False Positive Rate (FPR, number of false positives over the total number of negative descriptors) and True Negative Rate (TNR=1-FPR), where man-made structure is the positive class. The FPR, TPR and TNR values vary with the threshold applied to each model's likelihood ratio (Eq. 1).

Parameter setting: Following results reported in [14], where similar latent aspect modeling experiments are conducted, all our results are reported with a vocabulary size of 1000 visterms, and 60 aspects in aspect model 1 and 2.

4.2 Results

Fig. 4a displays the Receiver Operating Curve (ROC, TPR vs. FPR) of the two aspect models and the empirical distribution (baseline). As can be seen, the aspect model 1 performs slightly better than the empirical distribution method (although not significaly), while aspect model 2 outperforms the two other methods significantly, according to the paired T-test with a 95% confidence level.

To further validate our approach, Table 1 reports the Half-Total-Recognition Rate (HTRR) measured by 10-fold cross-validation. For each of the folds, 90% of the test data $\mathcal{D}_{test}$ is used to estimate the likelihood threshold T_{EER} leading to Equal Error Rate (EER, obtained when TPR=TNR) on this data. This threshold is then applied on the remaining 10% (unseen images) of $\mathcal{D}_{test}$, from which the HTRR (HTRR=(TPR+TNR)/2) is computed. This table shows that the ranking observed on the ROC curve is clearly maintained, and that aspect model 2 results in a 7.5% performance relative increase w.r.t. the baseline approach.

As mentioned in Section 3.2, aspect model 1 and the empirical distribution method assign specific visterms to the man-made or natural class independently of the individual documents in which those visterms occur. This sets a common limit on the maximum performance of both systems, which is referred here as the *ideal case*.

This limit is given by attributing to each visterm the class label corresponding to the class in which that visterm occurs the most in the *test data*. On our data, this *ideal case* provides an HTRR of 71.0%, showing that the visterm class attribution from empirical distribution and aspect model 1 is already close to the best achievable performance. Indeed, the class conditional probabilities shown in Figure 4b indicate that there is a substantial amount of polysemy. The class conditional probabilities are obtained by dividing the number of visterm occurrences in one class by the number of that visterm occurrences in both classes. Polysemy is indicated by the simultaneously quite high probabilities in both classes (e.g. for instance note that all visterms appear at least 15% in the *natural* class). Thus, in order to have a chance of performing better than the *ideal case*, visterms must be labeled differently depending on the specific image that is being segmented. This is the case with the aspect model 2 which, due to its ability to address the polysemy and synonymy ambiguities, is able to outperform the *ideal case*. More precisely, aspect model 2 switches visterm class labels according to the contextual information gathered through the identification of image-specific latent aspects. Indeed, in our data, successful class label switching occurs at least once for 727 out of the 1000 visterms of our vocabulary.

The impact of the contextual model can also be observed on individual images. Fig. 5 displays examples of man-made structure segmentation at T_{EER}. As we can observe in those images, aspect

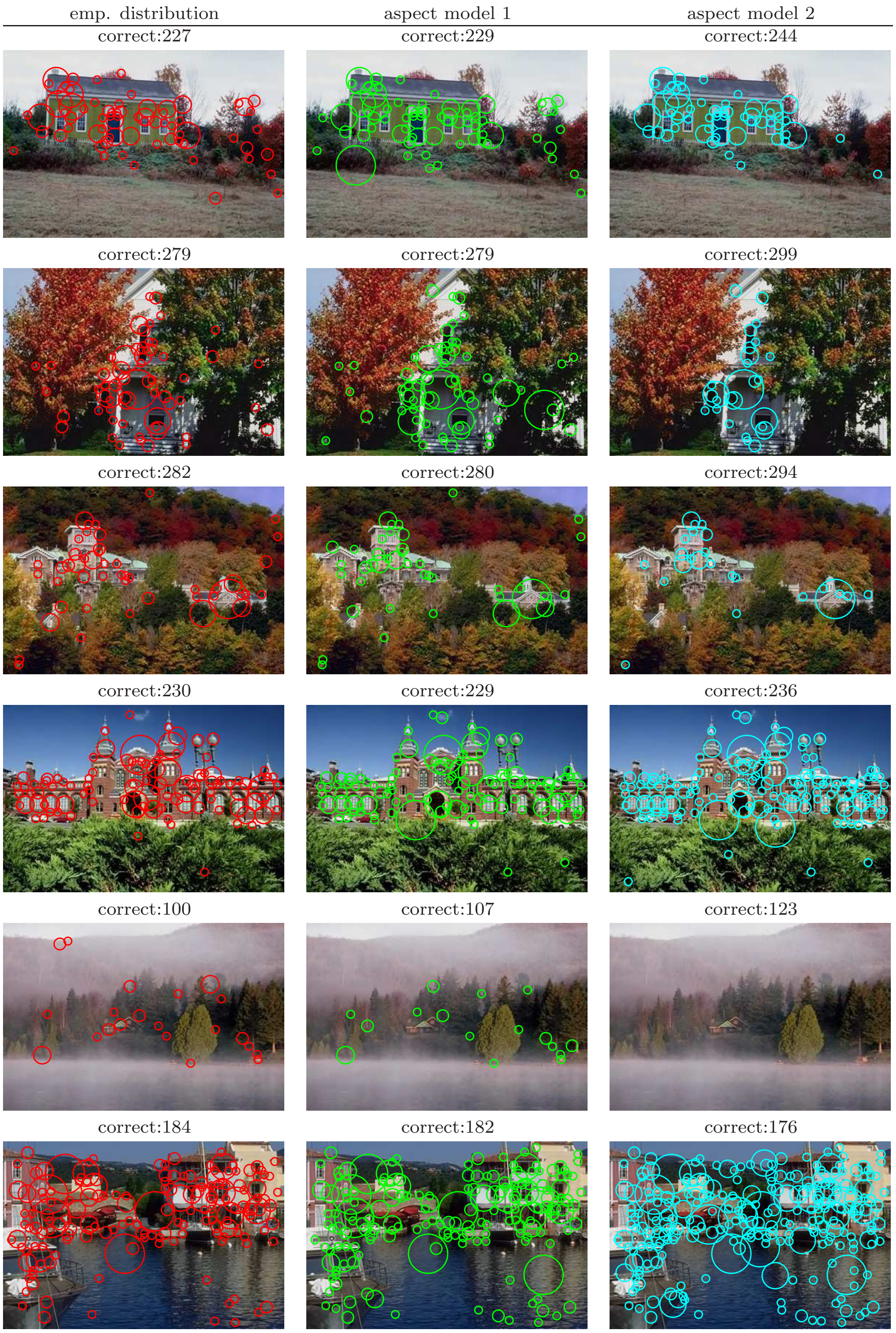

Figure 5: Image segmentation examples at T_{EER}. Results provided by: first column, empirical distribution; second column, aspect model 1; third column, aspect model 2. The total number of correctly classified regions (man-made + natural) is given.

model 2 improves the segmentation with respect to the two other methods in two ways. On one hand, in the first three examples, aspect model 2 increases the precision of the man-made segmentation, producing a slight decrease in the corresponding recall (some points in the man-made areas are lost). On the other hand, the fourth example shows aspect model 2 producing a higher recall of man-made visterms while maintaining a stable precision. In the fifth example, the occurrence of a strong context causes the whole image to be taken as natural scene. In the sixth example, however, the overestimation of the man-made related aspects leads to visterms that are dominantly classified as man-made. Nevertheless, overall, as indicated in Fig. 4 and Table 1, the introduction of context by co-occurrence is beneficial.

4.3 Markov Random Field (MRF) regularization

The contextual modeling with latent aspects that we present in this paper can be conveniently integrated with traditional spatial regularization schemes. To investigate this, we present the embedding of our contextual model within the MRF framework [4], though other schemes could be similarly employed [7,8].

Problem formulation. Let us denote by S the set of sites s, and by $\mathcal{Q}$ the set of cliques of two elements associated with a second-order neighborhood system $\mathcal{G}$ defined over S. The segmentation can be classically formulated using the Maximum A Posteriori (MAP) criterion as the estimation of the label field $C = \{c_s, s \in S\}$ which is most likely to have produced the observation field $V = \{v_s, s \in S\}$. In our case, the set of sites is given by the set of interest points, the observations v_s take their value in the set of visterms $\mathcal{V}$, and the labels c_s belong to the class set $\{man-made, natural\}$. Assuming that the observations are conditionally independent given the label field (i.e. $P(V|C) = \prod_s P(v_s|c_s)$), and that the label field is an MRF over the graph $(S, \mathcal{G})$, then due to the equivalence between MRF and Gibbs distribution ($P(x) = \frac{1}{Z}e^{-U(x)}$), the MAP formulation is equivalent to minimizing an energy function [4]

$$\begin{aligned} U(C,V) \quad = \quad & \underbrace{\sum_{s\in S} V_1(c_s) + \sum_{\{t,r\}\in\mathcal{Q}} V_1'(c_t,c_r)}_{U_1(C)} \\ & + \underbrace{\sum_{s\in S} V_2(v_s,c_s)}_{U_2(C,V)}, \end{aligned} \tag{11}$$

where U_1 is the regularization term which accounts for the prior spatial properties (homogeneity) of the label field, whose local potentials are defined by:

$$\begin{aligned} V_1(\text{man-made}) = \beta_p \;\; \text{and} \;\; V_1(\text{natural}) = 0, \\ V_1'(c_t,c_r) = \begin{cases} \beta_d & \text{if } c_t \neq c_r, \\ 0 & \text{otherwise,} \end{cases} \end{aligned} \tag{12}$$

β_d is the cost of having neighbors with different labels, while β_p is a potential that will favor the man-made class label (if $\beta_p < 0$) or the natural one (if $\beta_p > 0$). and U_2 is the data-driven term for which the local potential are defined by:

$$V_2(v_s, c_s) = -\log(p(v_s|c_s)). \tag{13}$$

To implement the above regularization scheme, we need to specify a neighborhood system. Several alternatives could be employed, exploiting for instance the scale of the invariant detector (e.g. see [8]). Here we used a simpler scheme: two points t and r are defined to be neighbors if r is one of the $N_{\mathcal{N}}$ nearest neighbors of t, and vice-versa. For this set of experiments we defined the neighborhood to be constituted by the five nearest neighbors. Finally, in the experiments, the minization of the energy function of Eq. 11 was conducted using simulated annealing [9].

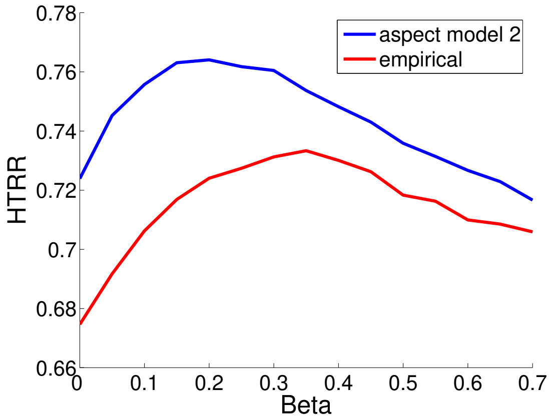

Figure 6: Half Total Recognition Rate for different β_d values.

Results. We investigate the impact of the regularization on the segmentation. The level of regularization is defined by β_d (a larger value implies a larger effect). The regularization is conducted by starting at the Equal Error Rate point, as defined in the 10-fold cross-validation experiments described in preceding Section. More precisely, for each of the folds, the threshold T_{EER} is used to set the prior on the labels by setting $\beta_p = -\log(T_{EER})$. Thus, in the experiments, when $\beta_d = 0$ (i.e. no spatial regularization is enforced), we obtain the same results as in Table 1. In Figure 6 we see that the best segmentation performance corresponds to an HTRR of 73.1% and a β_d of 0.35 with the empirical modeling, and an HTTR of 76.3% for a β_d of 0.2 and aspect model 2. This latter value of β_d is chosen for all the MRF illustrations reported in Figure 7 and 8.

The inclusion of the MRF relaxation boosted the performance of both aspect model 2 and empirical distribution. The MRF regularization improvement is higher for the empirical distribution model than for aspect model 2, but aspect model 2 still outperforms the empirical distribution when the best performance of each model are compared. This difference in improvement was to be expected, as aspect model 2 is already capturing some of the contextual information that the spatial regularization can provide (notice also that the maximum is achieved for a smaller value of β_d in aspect model 2).

Besides obtaining an increase of the HTRR value, we can visually notice a better spatial coherence of the segmentation, as can be seen in Figure 7 and 8. The MRF relaxation process reduces the occurrence of isolated points, and tends to increase the density of points within segmented regions. We show on the last row of Figure 7 that as can be expected when using prior modeling, on certain occasions the MRF step can over-regularize the segmentation, causing the attribution of only one label to the whole image.

5 Conclusion and future work

In this paper, we proposed computational models to perform contextual segmentation of images. These models enable us to exploit a different form of visual context, based on the co-occurrence analysis of visterms in the whole image rather than on the more traditional spatial relationships. Visterm co-occurrence is summarized into aspects models, whose relevance is estimated for any new image, and used to evaluate class-dependent visterm likelihoods. These models have been tested and validated on a man-made vs. natural scene image segmentation task. One model has clearly shown to help in disambiguating polysemic visterms based on the context they appear in. Producing satisfactory segmentation results, it outperforms a state-of-the-art likelihood ratio method. Moreover, we investigated the use of Markov Random Field models to introduce spatial coherence in the final

Figure 7: Effect of the MRF regularization on the man-made structure segmentation. The first two rows illustrate the benefit of the MRF regularization where wrongly classified isolated points are removed. The last row shows the deletion of all man-made classified regions when natural regions dominate the scene.

segmentation and show that the two types of context models can be integrated successfully. This additional information enables to overcome some visterm classification errors from the likelihood ratio and aspect models methods, increasing the final segmentation performance.

While the results presented here are encouraging, this task is complex, and there is a need for further improvements. Logical extensions would be the introduction of other sources of contextual information like color or scale and other forms of integration of spatial contextual information.

Acknowledgments

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 Data Mining (MULTI) project. Both projects are managed the Swiss National Science Foundation on behalf of the federal authorities.

References

[1] D. Blei, Y. Andrew, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1020, 2003.

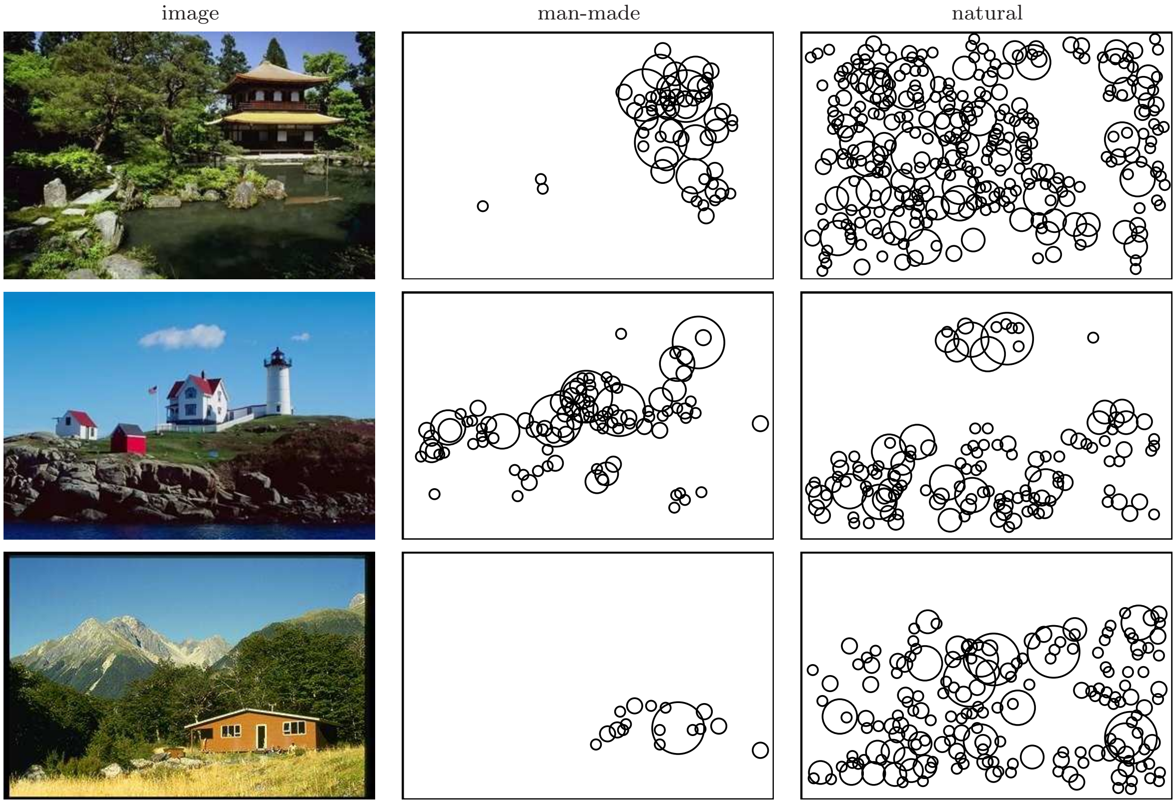

Figure 8: Illustration of the final segmentation with aspect model 2 and MRF regularization. This display avoids image clutter.

[2] G. Dorko and C. Schmid. Selection of scale invariant parts for object class recognition. In *Proc. of IEEE Int. Conf. on Computer Vision*, Nice, Oct. 2003.

[3] L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *Proc .of. IEEE Int. Conf. on Computer Vision And Pattern Recognition*, San Diego, Jun. 2005.

[4] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[5] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42:177–196, 2001.

[6] S. Kumar and M. Herbert. Discriminative random fields: A discriminative framework for contextual interaction in classification. In *Proc. of IEEE Int. Conf. on Computer Vision*, Nice, Oct. 2003.

[7] S. Kumar and M. Herbert. Man-made structure detection in natural images using a causal multiscale random field. In *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition*, Toronto, Jun. 2003.

[8] S. Lazebnik, C. Schmid, and J. Ponce. Affine-invariant local descriptors and neighborhood statistics for texture recognition. In *Proc. of IEEE Int. Conf. on Computer Vision*, Nice, Oct. 2003.

[9] S. Z. Li. *Markov Random Field Modeling in Computer Vision*. Springer, 1995.

[10] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[11] J. Malik, S. Belongie, T. Leung, and J. Shi. Contour and texture analysis for image segmentation. *International Journal of Computer Vision*, 43:7–27, 2001.

[12] K. Mikolajczyk and C. Schmid. Scale and affine interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.

[13] K. Murphy, A. Torralba, and W. Freeman. Using the forest to see the trees: A graphical model relating features, objects and scenes. In *Proc. of Neural Information Processing Systems*, Vancouver, Dec. 2003.

[14] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. V. Gool. Modeling scenes with local descriptors and latent aspects. In *Proc. of IEEE Int. Conf. on Computer Vision*, Beijing, Oct. 2005.

[15] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering object categories in image collections. In *Proc. of IEEE Int. Conf. on Computer Vision*, Beijing, Oct. 2005.

[16] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proc. of IEEE Int. Conf. on Computer Vision*, Nice, Oct. 2003.

[17] T. Tuytelaars and L. Van Gool. Content-based image retrieval based on local affinely invariant regions. In *Proc. of Int. Conf. on Visual Information Systems*, Amsterdam, Jun. 1999.

[18] J. Vogel and B. Schiele. Natural scene retrieval based on a semantic modeling step. In *Proc. of Int. Conf. on Image and Video Retrieval*, Dublin, Jul. 2004.

[19] J. Vogel and B. Schiele. A semantic typicality measure for natural scene categorization. In *Proc. of Pattern Recognition Symposium DAGM'04*, Tübingen, Sep. 2004.

[20] J. Willamowski, D. Arregui, G. Csurka, C.R. Dance, and L. Fan. Categorizing nine visual classes using local appearance descriptors. In *Proc. of LAVS Workshop, in ICPR'04*, Cambridge, Aug. 2004.