Contextual Analysis of Textured Scene Images

Markus Turtinen and Matti Pietikäinen
Machine Vision Group, Infotech Oulu
P.O.Box 4500, FI-90014 University of Oulu, Finland
{dillian,mkp}@ee.oulu.fi

Abstract

Classifying image regions into one of several pre-defined semantic categories is a typical image understanding problem. Different image regions and object types might have very similar color or texture characteristics making it difficult to categorize them. Without contextual information it is often impossible to find reasonable semantic labeling for outdoor images. In this paper, we combine an efficient SVM-based local classifier with the conditional random field framework to incorporate spatial context information to the classification. The images are represented with powerful local texture features. Then a discriminative multiclass model for finding good labeling for the image is learned. The performance of the method was evaluated with two different datasets. The approach was also shown to be useful in more general image retrieval and annotation tasks based on classification.

1 Introduction

Building vision systems for outdoor scene understanding has been a topic of wide interest in computer vision research [1]. For a human, scene understanding is relatively easy: one locates different object categories and their relative positions in the scene, and then perceives a description of the scene. Also in automatic analysis, different objects need to be detected and recognized. It is clear that different object detectors and recognition methods are the key elements in automatic analysis. It is also obvious that proper contextual information is important in automatic scene understanding. Consider, for example, an image region of blue sky or calm water. Their color characteristics can be very similar. By incorporating context information it is possible to improve the performance of individual object detectors and produce a more reasonable description of the scene.

A typical approach for scene classification is first to segment the image and then find the class labels for each region using a classifier trained to recognize different materials. Regions are usually represented with different color and texture features, and then a supervised learner is applied to the feature data. In this paper, we consider texture-based image classification and scene description by classifying image blocks into predefined categories [14]. Texture is attractive for use in outdoor image analysis. It is more robust than color with respect to changes in illumination, and it can also be utilized in night vision [2].

Powerful features and the local classification model form the basis for a good patch- or block-based scene labeling method. In addition, it is very important to take into account the context of the other regions because of the spatial dependency of the labels [5, 7, 19].

We propose a novel multiclass texture classification approach that can be used for segmenting textured scenes in a meaningful way. Texture information is extracted from the local image patches using an efficient multiscale and rotation invariant approach based on local binary pattern features [13]. The spatial regularity of the labels and the feature data are utilized in the contextual classification framework. The classification is performed by applying a multiclass SVM [20] on local image patches. Finally, a discriminative graphical model is learned to combine local classification with its context. The conditional random field (CRF) [9] is used for that purpose.

The proposed approach is a new kind of discriminative texture recognition system with many potential applications in computer vision. We demonstrate that the method can accurately predict labels, such as vegetation or a road, for different image regions in outdoor scenes. We also build a more general image annotation [4] and retrieval system over the contextual classification. This can be considered as an example application for managing personal digital image albums.

2 Related Work

Scene understanding and labeling in terms of identifying regions of interest in images is a traditional computer vision problem. Typically, the methods are based on color features extracted from the images. In [18], Saber et al. detected and recognized regions like sky and grass using color information modeled as 2D Gaussian functions. Texture information is often combined with color to improve the accuracy, as in [5] where gray level difference texture features were used with color. Pure texture statistics have been rarely used for classifying regions in outdoor scene images [2, 15].

Researchers of computer vision are now more frequently looking beyond low-level features and are interested in contextual issues [5, 19, 7, 8]. Two types of context are important: label context and data context. Label context means surrounding labels of the region of interest (i.e. sky label above tree label). Data context means the extracted data, such as pixels or features, from the neighboring regions. Both contexts are important because there is correlation between neighboring labels and image features. In addition to the label and data contexts, one can divide contexts broadly into local and global categories. The local context can be, for example, the local smoothness of pixel labels and the global context might be interaction of larger image regions.

Relaxation labeling [17] based approaches have been widely used for modeling the context in computer vision. These algorithms seek global consistency of labeling using constraints between labels and have been able to improve labeling in many applications. A more recent approach presented by Singhal et al. [19] uses probabilistic spatial context models to provide labels sequentially to image regions.

Markov Random Fields (MRF) are very commonly used undirected graphical models in computer vision [11]. They provide a theoretical way of incorporating local contextual constraints in labeling problems. MRF models are generative where the joint probability, $p(x, y) = p(x|y)p(y)$, of the features x and corresponding labels y are learned. The posterior, $p(y|x)$, over the labels given the data is expressed using the Bayes' rule and the prior over the labels, $p(y)$, is modeled as a MRF.

One problem of the traditional MRF is that it does not allow us to use the observed data, like image features, when modeling interactions between labels. In scene image

analysis such information is important and might improve interaction models. Conditional Random Fields (CRF) [9] allow us to incorporate data-dependent interactions in the models. Another advantage of CRF compared to the traditional MRF is that they are discriminative models: it is not required to generate prior distributions over the labels and the posterior $p(y|x)$ is modeled directly.

At the beginning, the CRF was applied to natural language processing, but later it has been proposed for various image analysis problems also. He et al. [7] learned local and global label features from the set of training images and used CRF for pixelwise scene labeling. Kumar and Hebert [8] used grid-based CRF and image features for detecting image blocks containing man-made structures. Weinman et al. [21] used a CRF model for detecting sign blocks from the natural images. The closest work of the one presented in this paper is [10] where SVM is also used with the CRF framework. They are restricted to binary classification cases when trying to segment brain tumors in MR imagery. We extend this model to multiclass problems, use more efficient parameter learning and inference techniques, and apply the model to a texture-based outdoor scene image classification. We also show that the contextual classification of images can be useful in more general image annotation and retrieval tasks.

3 System Overview

Our image classification method is based on efficient local texture features extracted from local image patches. Then the image blocks are classified in a contextual manner. The following subsections describe the methods in detail.

3.1 Image Feature Extraction

We use local binary pattern [13] features for describing the local image patches. LBP is invariant to monotonic gray-scale variations and has performed very well in various texture analysis problems. We extract this feature in multiscale fashion as illustrated in Fig. 1. The original large scene image is first divided into non-overlapping blocks. Then the features are extracted from these blocks on three different scales: windows of three different sizes are centered at each block and features are extracted from these by first scaling the patches to the same size. The features themselves from each scaled patch are also calculated using three different neighborhood radii ($r=1,2,3$ pixels). The main difference to [13] is that the original local image patches are processed in multiple scales to incorporate more information about the texture.

Rotation invariance is obtained by using circular neighborhood sampling and selecting only rotation invariant patterns as features. In practice, rotation invariant local patterns are such LBP codes that are rotated to their minimum values around the pixels [13]. By combining multiscale image representation with rotation invariant features we can model the texture of local patches very efficiently.

3.2 Classification Model

A grid-based graph is the most common graph topology in computer vision applications. Our model is also a grid (lattice) where each node corresponds to the image block from

Figure 1: Multiscale LBP feature extraction.

which the features are extracted. The nodes are connected to their adjacent nodes, and we construct a graph $G = (S, E)$ with set of nodes S and edges E .

CRF is a graphical discriminative model which takes the form

$$p(y|x) = \frac{1}{Z(x)} \exp \left(\sum_{i \in S} A(y_i, x) + \sum_{i \in S} \sum_{j \in N_i} I(y_i, y_j, x) \right), \quad (1)$$

where A and I are potential functions, S is the set of nodes and N_i is the neighborhood of the node i . $Z(x)$ is an observation-dependent normalization function. Function A is called association potential that couples features (x) from an image region (node) with a label (y) for the region. Function I is an interaction potential that couples image features for pairs of regions and labels.

In multiclass case (Y class problem) the association potential function is defined as

$$A(y_i, x) = \sum_{k=1}^Y \delta(y_i = k) \log P(y_i = k|x), \quad (2)$$

where $\delta(y_i = k)$ is 1 if $y_i = k$ and 0 otherwise. P is the posterior probability of the local discriminative classifier. Typically classifiers that provide directly posterior probabilities are used, such as the logistic classifier in the work of Kumar and Hebert [8]. We use a multiclass SVM with an RBF kernel as the local discriminative classifier. The original SVM is a binary classifier and does not provide posterior information, but the distance to the decision hyperplane and the posterior probabilities need to be found indirectly. The multiclass problem is first decomposed into several binary problems using one-vs-one coding. Then, pairwise class probabilities, $r_{kl} \approx p(y = k | y = k \text{ or } l, x)$, are estimated from the training data. These are estimated using a modified version of Platt's [16] method:

$$r_{kl} \approx \frac{1}{1 + \exp(A \cdot \hat{f} + B)}, \quad (3)$$

where $\hat{f}$ are the decision values (distances to hyperplane) and parameters A and B are estimated with the maximum likelihood method using the known training data. After that, the approach presented by Wu et al. [22] is used to obtain p_k from r_{kl} 's by solving the optimization problem:

$$\min_p \sum_{k=1}^Y \sum_{l:l \neq k} (r_{lk}p_k - r_{kl}p_l)^2 \quad \text{subject to} \quad \sum_{k=1}^Y p_k = 1. \quad (4)$$

The interaction potential predicts how the labels at two sites interact given the observations. In a multiclass case, the interaction potential gets the form

$$I(y_i, y_j, x) = \sum_{k=1}^Y \sum_{l=1}^Y (v_{kl} \cdot \Phi_{ij}(x)) (\delta(y_i = k) \delta(y_j = l)), \quad (5)$$

where v_{kl} are model parameters and $\Phi_{ij}(x)$ are features for sites i and j . In scene image analysis, these features can be designed so that they encourage label continuity. Instead of using just the distance (Euclidean or some other measure) between sites i and j as a feature, we use a similar approach to [10]: $\Phi_{ij}(x) = (\max(F(x)) - |F_i(x) - F_j(x)|) / \max(F(x))$, where $F_i(x)$ are texture features extracted from the node i and $\max(F(x))$ are the maximum values for each feature (evaluated from the training set). It was found that these kinds of edge features perform better compared to the distance features in a scene labeling application.

3.3 Parameter Learning and Inference

For learning the model parameters, we assume a set of labeled training images is available, and learn parameters sequentially. First, the local SVM classifier is trained using a grid search to find RBF kernel parameters γ and C . Then, these are fixed and the interaction potential parameters are learned maximizing the log-likelihood function $L(\theta) = \sum_{m=1}^M \log P(y^m | x^m, \theta)$ over M training images. Maximization is made with the gradient based method. The derivative of the log-likelihood with respect to parameters v_{kl} can be written as,

$$\frac{\partial L}{\partial v_{kl}} \sum_{m=1}^M \sum_{i \in S^m} \sum_{j \in N_i} (\delta(y_i^m = k) \delta(y_j^m = l) - \langle \delta(y_i = k) \delta(y_j = l) \rangle) \Phi_{ij}(x), \quad (6)$$

where $\langle \cdot \rangle$ brackets denote expectations with respect to the current model distribution. In practice, the expectation cannot be computed exactly due to the exponential number of configurations of labels y . In this work, we approximate expectations with the pseudo-marginals returned by loopy belief propagation (BP) [6]. Here we use a limited memory second-order optimizer (L-BFGS) that is very efficient in CRF parameter learning. For avoiding overfit we use regularization where Gaussian prior for the parameters is assumed. This is done by subtracting the term $\frac{\|v\|^2}{2\sigma^2}$ from the likelihood function and $\frac{v}{\sigma^2}$ from its gradient. For inference, the sum-of-product version of loopy BP is used to find the maximum posterior maginal (MPM) estimates of the labels for each node.

4 Experiments

The method was experimented with natural outdoor images taken from the Outex texture database [12] and the Sowerby image database [3]. The Outex database has two natural

outdoor scene image sets: the first set contains a sequence of 22 images taken by a human walking in the park (Outex_ID_NS00001), and the second one contains 20 random snapshots taken outdoors (Outex_ID_NS00000). All the images are 2272*1704 pixels in size. The Sowerby dataset contains 104 outdoor images taken around the Bristol area. The size of each image is 553*768 pixels. In addition, a subset of the Corel image database was used to demonstrate the potential of the approach in image annotation and retrieval problems.

The Outex images have sparsely defined ground truth regions for five different classes (grass, tree, road, sky and man-made structures etc.). Instead of using these sparse regions we first manually re-labeled each pixel in the images into these five categories to incorporate more contextual constraints. The Sowerby images had already pixel-wise labels.

First we experimented with the local classifier only. We compared the SVM classification with a logistic classifier which is a typical discriminative classifier used with CRFs. As a training data, half of the sequentially taken Outex_NS_00001 images (alternate indices) were used. The images were first divided into 24*24 pixels blocks and then the features were extracted from these blocks. We used the multiscale features described in Section 3.1. Three windows (sizes 24*24, 48*48 and 96*96 pixels) were set over each block and the region under them was scaled to 65*65 pixels using bilinear interpolation. Multiresolution and rotation invariant LBP with three radii (1,2,3) were extracted from each window and these feature vectors were concatenated. As a result, the feature data dimensionality was 162. The testing data was extracted from the other half of the images. With the logistic classifier, the classification rate was 79%, while with SVM it was 82%. In addition, we also experimented how multiscale feature extraction performs compared to the features extracted using only one window size. The classification rate was about 20 percentage units worse (62% with both classifiers) compared to the multiscale approach. Based on these experiments, it can be said that the SVM classification combined to the multiscale LBP features is a relatively effective local classification method.

Next, we experimented how contextual classification with the SVM and CRF performs. Similar features as before were extracted from the images. Then anisotropic CRF parameters (separate parameters for vertical and horizontal edges) were learned using the maximum likelihood method. Table 1 shows the results obtained with the Outex images. The NS_00001_train images were used for training the model and also the classification rates for these training images are reported. The classification rate for the other half of the NS_00001 image set achieved 87.6 %. For comparison, we also classified the original sparsely defined ground truth regions (GT) from the Outex images. The classification rate was 95.2 %, which is almost 10 percent units better than the best result reported in [15] (85.4 %). The same model was also experimented with different images taken from the Outex_NS_00000 set. This image set is more challenging, but the classification rate achieved was 77.1 % without re-learning the model using images taken from this particular data set. For original sparse ground truth regions the accuracy was 82.5 %. Some example images and classification results are shown in Fig. 2.

The Sowerby image database has several labels in different categories. We selected six main categories for our experiments: man-made structures, vegetation, road, sky, road borders and undefined. Undefined regions were not considered when calculating the classification rate. Similar features to those in the Outex experiments were extracted from the images, except the block sizes were 8*8, 16*16 and 32*32 pixels. 60 images were randomly selected for training and the remaining 44 for testing. The classification rate

Image set	SVM [%]	SVM (GT) [%]	CRF [%]	CRF (GT) [%]
NS_00001_train	85.3	95.7	92.2	99.2
NS_00001_test	82.2	91.8	87.6	95.2
NS_00000_all	67.8	71.8	77.1	82.5

Table 1: Classification results with Outex images.

Figure 2: Example result images from Outex database.

with the local classifier was 63.9 % and the contextual classification result was 80.3 %. Fig. 3 shows two example images of classification results.

Reliable prediction of class labels of different regions in the image is important and useful in many different applications. We experimented how texture based contextual image classification performs in typical digital album management tasks, like automatic annotation of images and retrieval using keywords. We used a subset of the Corel image database in the experiments. The model was trained using sparsely labeled images with the following six labels: animal with fur (polar bear and bear), vegetation, water, sky, buildings (castle) and undefined. Then the model was applied to new kinds of images to predict the class labels for regions. After classification, a simple rule-based approach was applied to annotate and categorize images into nature, structured and uncategorized scenes. The nature category images contain vegetation, sky, water and sometimes also animals. The structured scenes have mostly buildings and sky, but may also contain some water and vegetation regions.

Annotation was started by calculating the class specific scores for each class. This was performed by calculating the sum of the posterior probabilities of final classification for each connected component larger than a specified threshold (10 blocks). To make the annotation more confident, only those images whose summed score of other than undefined regions was more than 40% of the maximum were annotated as a nature or

Figure 3: Example result images from Sowerby database.

structured scene. Otherwise the image was put into the uncategorized group. If there was some building score in these images, they were annotated as structured scenes. Otherwise, if there were vegetation or animals, the image was annotated as being a nature scene type. Totally 40 test images from 70 had labeled regions whose total score was more than 40% of the maximum. Using the local classifier alone, only seven images scored above that limit. In total 36 images of these 40 were put into the correct category. Using the local classifier alone, one image out of seven was wrongly categorized.

In the keyword based retrieval the scores were defined similarly. Then the images were ordered based on user query and scores. The query can be applied to all the images or only for those whose confidence level is above the limit (i.e. the total score is more than 40% of the maximum). Fig. 4 shows two example queries applied to all the images (above) and for those whose total score is above 40% of the maximum (below). Five images with the highest scores are shown in both cases.

Figure 4: Example queries using keywords animal+vegetation for all the images (above) and building+sky for those whose confidence is high (below).

5 Conclusions

In this paper, we proposed a method for contextual multiclass classification of textured images based on efficient local features, an SVM classifier and a CRF framework. Multiscale and rotation invariant local binary pattern texture features were extracted from

the image patches. Then local image classification with the support vector machine was combined with a conditional random field to include contextual information in the image classification.

The approach was experimented with outdoor scene images taken from two different databases. Even though the color information might be useful in these particular datasets, the classification rates with gray-scale images and texture features were very good. The contextual information improved the classification rate and the SVM was found to be more efficient with our features than the logistic classifier that is typically used in the CRF framework. The model is also easy to include in more general high-level image analysis applications. We demonstrated its usefulness in automatic image annotation and keyword based retrieval using images taken from the Corel data set.

In the future, we should pay more attention on the design of the edge features to control the spatial regularization in the texture classification. Instead of using a regular lattice graph, it would also be interesting to study how sparse regions detected by different interest point approaches could be used in this kind of image classification. Finally, it would be very interesting to build a real digital image album management system and study how to keep the models up-to-date and how to add new class detectors to the models.

Acknowledgments Financial support provided by the Infotech Oulu Graduate School and the Academy of Finland is gratefully acknowledged.

References

- [1] J. Batlle, A. Casals, J. Freixenet, and J. Marti. A review on strategies for recognizing natural objects in colour images of outdoor scenes. *Image and Vision Computing*, 18(6–7):515–530, May 1999.
- [2] R. Castano, R. Manduchi, and J. Fox. Classification experiments on real-world texture. In *Third Workshop on Empirical Evaluation Methods in Computer Vision*, pages 3–20, 2001.
- [3] D. Collins D, W. Wright, and P. Greenway. The Sowerby image database. In *International Conference on Image Processing and Its Applications*, pages 306–309, 1999.
- [4] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *European Conference on Computer Vision (ECCV)*, pages 97–112, 2002.
- [5] X. Feng, C. Williams, and S. Felderhof. Combining belief networks and neural networks for scene segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):467–483, April 2002.
- [6] B. Frey and D. MacKay. A revolution: Belief propagation in graphs with cycles. In *NIPS*, pages 479–485, 1997.
- [7] X. He, R. Zemel, and M. Carreira-Perpinan. Multiscale conditional random fields for image labeling. In *IEEE International Conference on Computer Vision and Pattern Recognition*, volume 2, pages 695–702, 2004.

- [8] S. Kumar and M. Hebert. Discriminative random field: A discriminative framework for contextual interaction in classification. In *IEEE International Conference on Computer Vision*, pages 1150–1157, 2003.
- [9] J. Lafferty, F. Pereira, and A. McCallum. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning*, pages 282–289, 2001.
- [10] C-H. Lee, R. Greiner, and M. Schmidt. Support vector random fields for spatial classification. In *9th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, pages 121–132, 2005.
- [11] S. Li. *Markov Random Field Modeling in Image Analysis*. Springer-Verlag, Tokyo, 2001.
- [12] T. Ojala, T. Mäenpää, M. Pietikäinen, J. Viertola, J. Kyllönen, and S. Huovinen. Outex - new framework for empirical evaluation of texture analysis algorithms. In *Proc. ICPR*, volume 1, pages 701–706, Quebec, Canada, 2002.
- [13] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.
- [14] R. Picard and T. Minka. Vision texture for annotation. *Multimedia Systems*, 3(1):3–14, 1995.
- [15] M. Pietikäinen, T. Nurmela, T. Mäenpää, and M. Turtinen. View-based recognition of real-world textures. *Pattern Recognition*, 37(2):313–323, 2004.
- [16] J. Platt. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74, 2000.
- [17] A. Rosenfeld, R. Hummel, and S. Zucker. Scene labeling by relaxation operations. *IEEE Transactions on System, Man, and Cybernetics*, 6:420–433, 1976.
- [18] E. Saber, A. Tekalp, R. Eschbach, and K. Knox. Automatic image annotation using adaptive color classification. *Graphical Models and Image Processing*, 58(2):115–126, 1996.
- [19] A. Singhal, J. Luo, and W. Zhu. Probabilistic spatial context models for scene content understanding. In *IEEE International Conference on Computer Vision and Pattern Recognition*, volume 1, pages 235–241, 2003.
- [20] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1996.
- [21] J. Weinman, A. Hanson, and A. McCallum. Sign detection in natural images with conditional random fields. In *IEEE International Workshop on Machine Learning for Signal Processing*, pages 549–558, 2004.
- [22] T-F. Wu, C-J. Lin, and R. Weng. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5:975–1005, 2004.