Semantic Segmentation of Street-Side Images

Michal Recky[1], Franz Leberl[2]

[1]Institute for Computer Graphics and Vision
Graz University of Technology
recky@icg.tugraz.at

[2]Institute for Computer Graphics and Vision
Graz University of Technology
leberl@icg.tugraz.at

Abstract
In this paper we propose a method for semantic segmentation of street-side images. Segmentation and classification is pixel based and results in classes of building facades, sections of sky, road and other areas present in general images taken in the urban environment. A segmentation method is suggested and detected segments are classified. Final classification is reinforced using context information implemented in a form of the discriminative random fields (DRF). Results show that this approach can overcome problems with the lack of features, as additional constraints are used in the classification.

Keywords: context, semantic segmentation, discriminative random fields

1 Introduction

Even rapid progress in object recognition has left the recognition task a challenging problem. The best algorithms today cannot compete with human vision. The possible reason for this is that in human vision, object recognition is a global process. In computer vision, most algorithms are focused on a specific object class and tend to neglect overall context information in the image. Background information is considered ineffective and gets removed. But in the human vision, background and contextual information play a major role in a recognition task. It is therefore suggested that the context is a basic element of a successful recognition algorithm [6, 14]. For example, the application of image context information during the recognition task can narrow a recognition area and thus eliminate false positives.

The goal of a semantic segmentation algorithm is to retrieve the image context information through classification of each region in the image into some predefined classes (usually definable for an application domain and type of image). In this paper, we consider single street-side images and introduce a semantic segmentation method for this specific domain. Common classes in street-side scenes are the building facades, sections of road (ground level) and sections of sky. We also consider vegetation, clouds, building roofs and grass areas. Dark areas in the image are marked in dark gray and are not classified. (see Figure 1).

Figure 1: A street-side image and the classification result, using the method proposed in this paper. Classes in the image are marked in different color.

At first, an image is segmented into large, logically coherent regions. It is assumed that only one object class is associated with one region. During the subsequent classification, only regions are considered as the objects of classification. To improve the result of classification, spatial relations between the segments are examined.
In this approach, not only the visual features of the segments are used for categorization, but it is assumed, that there are some spatial and contextual relations between the classes in street-side scenes. An example of spatial rules is the fact that in a majority of properly oriented street-side images, building facades are located below sections of the sky. If in the classification output this probability rule is not met, it may indicate an error in classification. These spatial rules get represented as a discriminative random field (DRF). Classifiers are learned in a supervised process. As it is described in [7], only a small number of training pictures is required to train the classifiers. For this purpose, we developed a hand-labeled ground truth database. The same database is used for training of the discriminative random fields [16].

2 Related work

Context understanding in an urban environment is gaining relevance due to initiatives such as Google Earth or Microsoft Live Maps. Understanding of context in 3D modeling is a well known idea. Different sources, from single image [3, 7] to video sequence [2], have been considered for context retrieval. In the single image case appearance based and shape based methods are used for this purpose. Features like color, texture, shape and some geometric cues [7] have proven to be reliable sources of information. In the case of a video sequence or multiple images of the same scene, 3D point clouds can be used to retrieve context [2].
Enhancing classification by exploiting spatial dependencies has been suggested by several authors. The Markov random fields model (MRF), as a representation of spatial dependencies, is a classical approach in this case [1, 5]. However, recent work shows that conditional random fields are an improvement over the MRF model in the labeling task [11, 17]. This is due to better discriminative properties of the conditional random fields over MRF [18]. Discriminative random fields are a specific application of conditional random fields and have been suggested for the task of categorization in the work of Kumar [10]. In this paper, we propose to apply the DRF to verify (and eventually select correct) classification of the image regions. We show that this method will provide better result in situations, where other classification methods reported problems with the lack of features [7].

3 Segmentation

Localization of the region borders and position is formulated as a segmentation problem. Several requirements must be met by this segmentation to cope with the problems presented in a street-side scenery. At first, regions have to be logically coherent. It means for example, a single region should contain only one building façade (or part of it) and should not extend to facades of different buildings, or to the sky or ground regions. But also, segmented regions should be as large as possible.

When examining pictures of street-side scenery, it is obvious that texture covariance can change rapidly through a logically coherent region. As an example, in one building façade, regions with low covariance alternate with regions containing ornaments or pillars, where covariance is high. But both of these regions may still be part of the same building façade, so we would like it to be considered as one region. This requirement is not easily met, because segmentations are usually designed to distinguish between such regions. Also, borders between two regions can be well-defined in street-side images, but they may also be very smooth (for example, between clouds and sky regions).

To meet these requirements, we use a non-standard segmentation approach with a novel variation. Watershed segmentation serves as a primary segment retrieval method. The threshold for the segmentation is set low so that the image gets intentionally over-segmented. In the next step, segments which are geometrically close to each other and are similar are joined into larger regions. Similarity of the segment is computed using color.

We define “Visual similarity” as a floating point value between 0 and 1 expressing how similar two color values look like (what is their visual difference). In most cases, visual similarity can be computed in the CIE-lab color space, as a Euclidean distance of lab values [12]. However, implementation revealed that this approach is not suitable in the current application. The main reason is that in CIE-lab space, hue and saturation have approximately the same weights in computing similarity. In street-side images, most building facades can be distinguished by their hue, but nearly all facades have a rather low saturation. Therefore, to differentiate between two buildings, a large weight must be put on hue, and smaller on saturation. To achieve this, visual similarity is computed through a specific formula in HSV color space:

$$\varphi(C_1, C_2) = |h_1 - h_2| \cdot \min\left(f_1(\max(s_1, s_2)),\ f_2(avg(b_1, b_2))\right) \quad (1)$$

Where $C_1 = [h_1, s_1, b_1]$ and $C_2 = [h_2, s_2, b_2]$ are colors in HSV color space and f_1, f_2 are logarithmic functions:

$$f_1(x) = \frac{1}{Z_1}\log(k_1 x + 1) \quad (2)$$

$$f_2(x) = \min\left(\frac{1}{Z_2}\log(k_2 x + 1), 1 - \left(\frac{1}{Z_3}\log(k_3 x + 1)\right)\right) \quad (3)$$

where Z_1, Z_2, Z_3 are normalizing constants (normalizing f_1 and f_2 into <0,1>).

Similar modifications are used for differences of saturation and brightness. A final visual similarity value is computed as maximum of the differences of hue, saturation and brightness. In this approach, several variable coefficients (k_1, k_2, k_3, …) are used (in logarithmic functions). To achieve

best results, these coefficients have been optimized in a supervised learning process. Hand-labeled validation dataset (with each building marked as different object) was used, and for each set of coefficients, segmentation was performed. Coefficients that achieved the best results are subsequently used in segmentation. In this approach, it is not necessary to compute transformations between CIE-lab and HSV color space and still compute similarity values with modifiable weights on hue, saturation and brightness. The logarithmic functions were chosen to simulate the requirements on HSV parameters, as these functions can narrowly modify weights when required (close to zero) and still remain nearly constant in higher values.

Merging of segments into regions is an iterative process. In the first step, only segments larger then 0.2% of the image and visually similar are merged into a composite region (more then two segments are allowed to merge in one step). Subsequently, smaller segments are merged into existing regions. Also, visual similarity is computed and required for merging, but the similarity threshold is reduced with each step. The representative color of the region is recomputed after each step. In this approach, it is assumed, that large segments are more important for the subsequent classification, as they are usually representing some coherent areas in the image. On the other hand, small segments may represent some small objects, or texture elements. Therefore, large segments have the priority in the merging step, but the requirements for their merging are high.

This approach for image segmentation has several advantages over the standard methods. As described before, segmentation can be easily modified by adjusting the coefficients, obtained from ground truth data. By over-segmenting the image in a watershed segmentation, most details are preserved, so in the final output, borders of the regions are well-defined. Also, when we proceed with merging of segments that are geometrically close to each other, but not necessary connected, regions in the final segmentation do not have to be continuous. This is especially useful in urban areas, where building facades or other logically coherent areas are often dissected by wires, traffic lamps, poles, or other objects in the image. These areas can then still be considered as uniform regions.

4 Classification

In a segmentation step, several large regions are usually detected in the image. These regions are subsequently classified into building facades, sky, cloud, roof, ground, vegetation and grass classes. Regions with intensity <0, 1> lower then 0.1 are marked as dark/unclassified, as in our database they lack any features necessary for the classification (due to camera quality). Only regions larger then 1% of the image are classified. Labeling of smaller regions is decided based on their neighborhood segments.

Classification is based on a decision tree. Training is a supervised process. As a result of accurate segmentation, standard open ground truth databases (like LabelMe) are not precise enough. Therefore, a new ground truth dataset was created and each region was manually classified. We use 30 hand labeled images as a training set, 200 images remain for testing purposes.

In the process of classification, each region is considered a coherent object. Classification is based on color, position in the image, size and texture. A single representative color value is computed for a region as an average of color of the pixels inside the region. In a learning process, a color histogram is created for each category. In a classification process, the color of the region is compared with a class histogram.

The position of the region in an image is represented as a position matrix. The image gets divided into a regular grid; each cell in the grid represents a coefficient in a matrix. It is computed if the region belongs to the cell. The same process is applied during the position classifier training. In the classification step, the position matrix of the region and the position matrix of the class are compared.

The texture of the region gets expressed as a histogram of gradient values over the region area. This representation of the texture is sufficient to distinguish smooth regions from textured regions. In the process of image over-segmentation, textured areas get segmented. As described in the previous section, these areas may be subsequently joined, so the insides of the regions may contain high gradient values. Therefore, classes like building roofs, or vegetation areas that contain some texture information relevant for classification, can be recognized thanks to this feature.
For classification, we use a decision tree. The last level of the tree contain the confidence values for each class computed as a joint probability of the classifiers located in the path from the root to the leaves. These values may be considered as the classification result, but as described in [7], features presented in this section may not be sufficient to discriminate between all classes. For example, regions of the sky and regions of façade windows can be very similar in color and texture and they may be located in similar positions in the image. Therefore, it is necessary to use some additional constraints in the classification. In the next section, we present spatial rules for verification of the classification.

5 Spatial rules in classification

Real objects in street-side scenes are in specific spatial relations to each other. For example, sky and clouds are always above the buildings, roofs are usually above the facades and ground is below the buildings. It is assumed that some of these rules are transferable into digital images as a central projection of the real scene. Using these rules may be valuable as constraints in the classification. Before applying the rules, we must assume that we already know what objects may be located in the image. Therefore, in the process of classification, not only the most likely classification result is used, but several results with highest confidence value are used as a set of competing classification hypotheses (see Figure 2).

Figure 2: Multiple hypotheses for a single image. For each region, all possible classifications passing the threshold are included in the hypotheses. Note the variation in roof/façade and cloud/façade classification in different hypotheses.

For each hypothesis, spatial rules get checked. Finally, a winning hypothesis is selected, as the one for which the maximal number of spatial rules is valid.
Spatial rules are encoded as a probability of spatial relations between two different classes. To extract the spatial rules that are commonly valid in street-side images, we must have a labeled ground truth database, with all objects classified.
Spatial rules are implemented as a discriminative random field data structure, representing every region in the image as a graph node. Regions close to one another are neighbors in the graph. In a classical approach, where each node represents a pixel, or a grid element in the image, spatial relations are implicit in the position of such element in the picture [9]. In our model, regions are not

assembled in any predictable fashion and they vary in shape and size. To extract the spatial relation, the graph structure is assembled with the image regions as the nodes (see Figure 3).

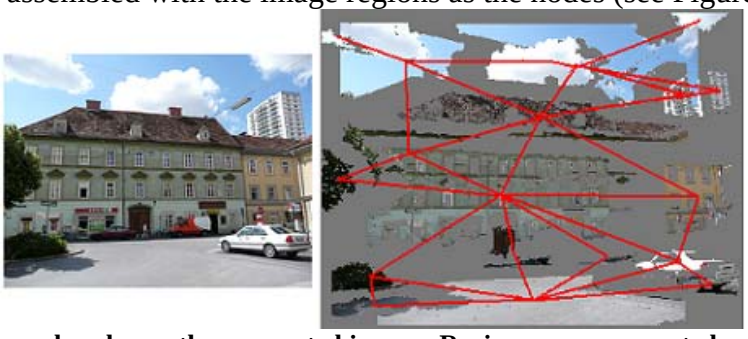

Figure 3: Graph-structure placed over the segmented image. Regions are represented as a graph node. An edge is placed between each two neighboring region. This graph is the basic data structure for DRF.

In the case of street-side images, mostly vertical spatial relations are relevant. Relations that are examined between the regions are described in Table 1.

Region i	Relation	Region j
Bounding box	*Inside* *Enveloping* *partially above* *partially below* *fully above* *fully below*	*Bounding box*
Region centre	*Inside* *Above* *Below*	*Bounding box*
Region centre	*Above* *Below*	*Region centre*

Table 1: Spatial relations are described based on the relations between bounding boxes and centers of two regions.

.

Let us assume we want to represent the conditional distribution $P(\mathbf{x}|\mathbf{y})$ of spatial relations over classes ($\mathbf{x}$ is a vector representing classes and $\mathbf{y}$ are the observations). According to the Hammersley-Clifford theorem [5], this distribution can be expressed as

$$P(\mathbf{x} \mid \mathbf{y}) = \frac{1}{Z} \exp\left(\sum_{i \in S} \mathrm{A}_i(x_i, \mathbf{y}) + \sum_{i \in S} \sum_{j \in N_i} \mathrm{I}_{ij}(x_i, x_j, \mathbf{y}) \right) \quad (4)$$

where Z is normalizing constant, S is the set of nodes, N_i is the set of neighbors of node $i \in S$. A_i is the unary (association) potential and I_{ij} is the pairwise (interaction) potential. In this form of distribution, the assumption of acyclicity is no longer required, which is a certain advantage in our application.

The unary potential A_i represents the measure of how likely node i belongs to class x_i, given the observation vector $\mathbf{y}$. In our approach, this potential is directly derived from the previous classification step. In classification, confidence values of each class were assigned to each region. These values, when normalized, serve as the unary potentials.

Pairwise potentials I_{ij} represents the measure of interaction between two neighboring nodes i and j given the observation vector $\mathbf{y}$. Pairwise potentials are derived from the training set during the learning process. Let us assume, that M is the set of training images, $\mathbf{x}_k$ is the classification of k-th image and $\mathbf{y}_k$ is observation of spatial relation in k-th image (see Table 1). We can represent the set

of classified regions neighboring the region i in image $k \in \mathrm{M}$ as $\mathbf{x}_k^{N_i}$. Then the probability that region i in the k-th image is classified into class x_i is $P\left(x_{ik} \mid \mathbf{x}_k^{N_i}, \mathbf{y}_k\right)$. This value can be computed directly from the training set. Inserting this value into equation (4) gives us the parameters for pairwise potential I_{ij}, as described in [9]

6 Results

For testing purposes, 230 images with different weather and lighting conditions were selected from the ICG Vienna and Graz database. These images contain a large variety of objects from historical buildings, standard city blocks, residential apartments and modern architecture (see Figure 4).

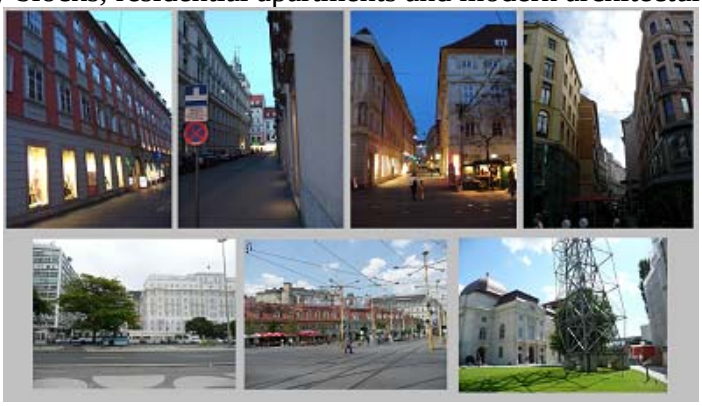

Figure 4: Example of the urban scenes present in the database. Note the variation in illumination, view positions and architectural objects in the images.

To speed up and normalize the testing process, images were down-sampled to 0.3 Mpix resolution (640x480). Thirty of these images were used in a supervised training process as the hand-labeled ground truth data. Tests were performed with a following hardware setup: Intel 2660Mhz, 2 GB RAM, GeForce 8800. In this configuration, segmentation and classification of the image takes approximately 2 seconds.

As a first experiment, we demonstrate the improved segmentation performance by using the novel visual similarity calculation. To test the performance of segmentation, 50 testing images were selected. In this set of images, each building façade was labeled differently. The segmentation of images, based on visual similarity and CIE-lab distance was computed. For each image and each building façade, the area of the façade region extending to other than original coherent area was computed. In the case of CIE-lab distance, this was approximately 5.7% of the region (average of all façade regions in all testing images). In case of visual similarity, this area was reduced to 3.2%.

	facade [%]	**roof [%]**	**ground [%]**	**sky [%]**	**vegetation [%]**	**grass [%]**	**cloud [%]**
clas	*89,3*	*76,5*	*92,4*	*97,6*	*80,4*	*93,5*	*57,5*
with DRF	*93,7*	*85.2*	*94,3*	*98,1*	*83,7*	*95,4*	*62,3*

Table 2: Results of the classification. In the first row, only visual features were used for the classification. In the second row classification was reinforced by discriminative random fields.

In Table 2 we can see correct classification rates for each class in the testing database. The percentage numbers express the value of correctly classified pixels of each class presented in the image. When computing the average over all testing images, contribution of each image was weighted by a size of area covered by a class. Regions with intensity lower then 0.1 were marked as dark areas. When present in specific areas of the class, these regions were not considered as

incorrectly classified. Classification is performing worst in the cloud area. This is due to the weak visual differences between the sky and clouds and difficult segmentation of the area. Using DRF for verification provided best results in roof and façade areas, as these have strong contextual relations to other classes, but their appearance-based classification is bad.

7 Conclusion

In this paper, a method for street-side semantic segmentation is presented. This method is able to solve problems of previous classification approaches (when visual features are too similar for two classes). This is due to the new verification method based on spatial relations between the classes applied to the large regions of the image. Seven different classes can be detected (see Figure 5), but this number can be further increased, when an extended ground-truth database will be available. Also, a novel method of segmentation that can be tailored (in automatic supervised learning process) to a specific image domain is presented as an improvement over the standard segmentation approaches.

Acknowledgments: This work has been supported by the Austrian Science Fund (FWF) under the doctoral program Confluence of Vision and Graphics W1209

References

[1] ALTUN Y., HOFMANN T., SMOLA A.; Hidden markov support vector machines. *20th ICML*, 2003

[2] BROSTOW G., SHOTTON J., FAUQUEUR J., and CIPOLLA R.; Segmentation and recognition using structure from motion point clouds. *European Conference on Computer Vision (ECCV)*, pages 44–57. 2008.

[3] CARBONETTO P., FREITAS N., BARNARD K.; A statistical model for general contextual object recognition. *European Conference on Computer Vision (ECCV)*, pages 350–362, 2004.

[4] FELZENSZWALB P.F., HUTTENLOCHER D. P.; Efficient graph-based image segmentation. *Int. J. Comput. Vision*, 59(2):167–181, September 2004.

[5] HAMMERSLEY J.M., CLIFFORD P.; Markov field on finite graph and lattices. *Unpublished*

[6] HEITZ G, KOLLER D.; Learning Spatial Context: Using Stuff to Find Things. *European Conference on Computer Vision (ECCV)*, 2008

[7] HOEIM D., EFROS A. A. and HEBERT M.; Geometric context from a single image. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 654–661 Vol. 1, 2005.

[8] KONISHI S., YUILLE A.L.; Statistical cues for domain specific image segmentation with performance analysis. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 1, 2000.

[9] KUMAR S.; Models for Learning Spatial Interactions in Natural Images for Context-Based Classification. *Carnegie Mellon University*, Pittsburg 2005

[10] KUMAR. S, HERBERT M.; Discriminative random fields. *International Journal of Computer Vision*, 68(2):179–201, June 2006

[11] LAFFERTY J., PEREIRA F.; Conditional random fields: Probabilistic models for segmenting and labeling sequence data, 2001.

[12] MELGOSA M.; Testing CIELAB-based color-difference formulas. *Color Research & Application Volume 25 Issue 1*, pages 49 – 55

[13] OHTA Y.; Knowledge-based Interpretation of Outdoor Natural Color Scenes. *Pitman, 1985*

[14] RABINOVICH A., VEDALDI A., GALLEGUILLOS C., WIEWIORA E., BELONGIE S.; Objects in context. *ICCV*, 2007

[15] SWAIN M.J., BALLARD D.H.; Color indexing. *International Journal of Computer Vision*, 7(1):11–32, November 1991.

[16] WALLACH H.; Efficient training of conditional random fields. Master's thesis, University of Edinburgh, 2002.

[17] XUMING HE, ZEMEL R. S. and CAREIRA-PERPINAN M. A.; Multiscale conditional random fields for image labeling. In *Computer Vision and Pattern Recognition*, volume 2, pages II–695–II–702 Vol.2, 2004.

[18] ZHONG P., WANG R.; Object detection based on combination of conditional random field and markov random field. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 3, pages 160–163, 2006.

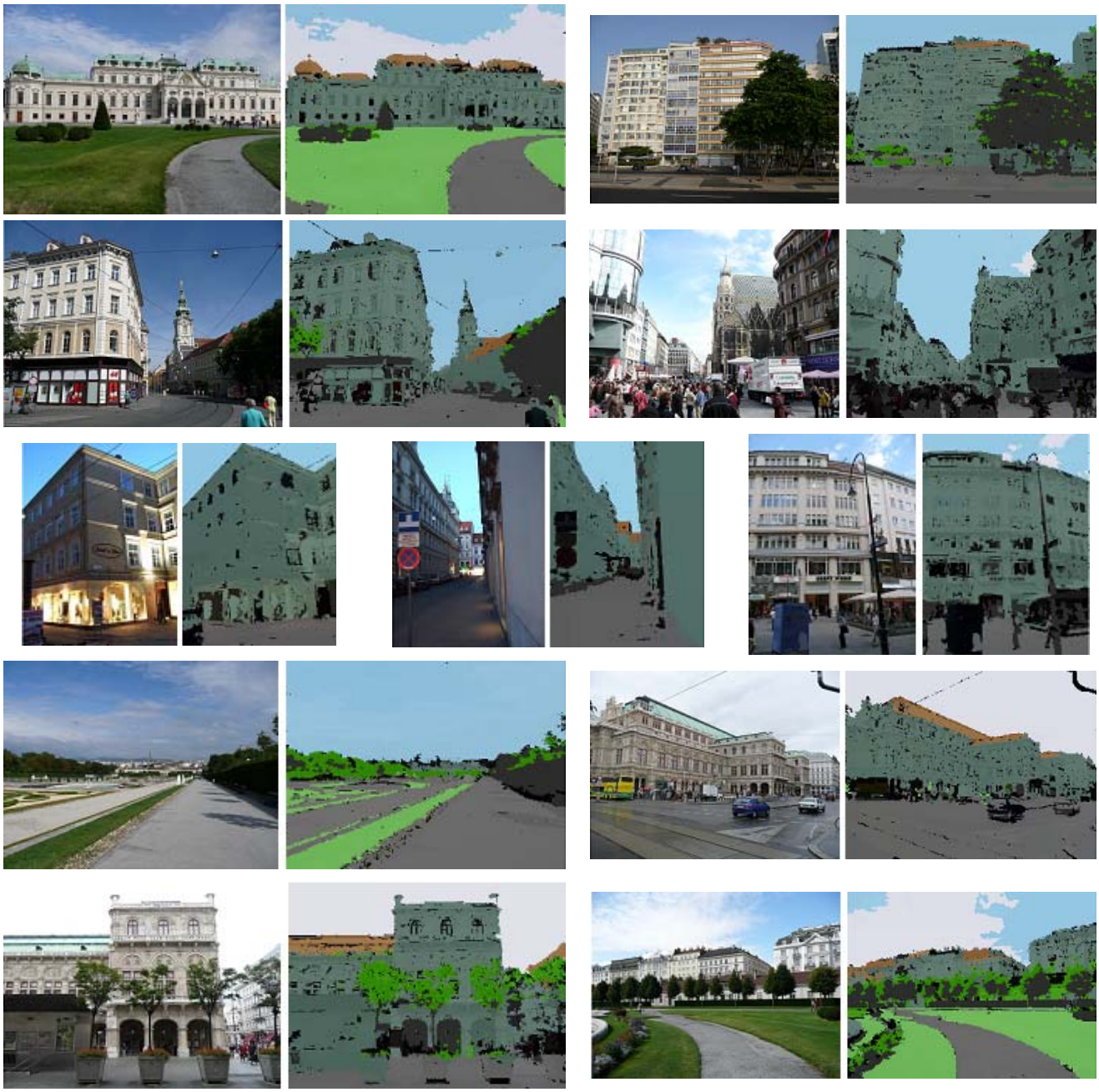

Figure 5: Examples of image context. Each class is marked in different color. Dark green – building façade, brown – roof, gray – ground, green – vegetation, blue – sky, light green – grass, dark gray – shadow, black - unclassified. We can see in these examples that most problems are in false positives for the vegetation class and false negatives for cloud class. Unidentified areas are mostly small, visually distinctive regions, like windows, or pedestrians.