A Bayesian Approach for Scene Interpretation with Integrated Hierarchical Structure

Martin Drauschke[1] and Wolfgang Förstner[2]

[1] Institute of Applied Computer Science, Bundeswehr University Munich, Germany
[2] Institute of Geodesy and Geoinformation, University of Bonn, Germany
martin.drauschke@unibw.de, wf@ipb.uni-bonn.de

Abstract. We propose a concept for scene interpretation with integrated hierarchical structure. This hierarchical structure is used to detect mereological relations between complex objects as buildings and their parts, e. g., windows. We start with segmenting regions at many scales, arranging them in a hierarchy, and classifying them by a common classifier. Then, we use the hierarchy graph of regions to construct a conditional Bayesian network, where the probabilities of class occurrences in the hierarchy are used to improve the classification results of the segmented regions in various scales. The interpreted regions can be used to derive a consistent scene representation, and they can be used as object detectors as well. We show that our framework is able to learn models for several objects, such that we can reliably detect instances of them in other images.

1 Introduction

Scene interpretation is a very active research field in computer vision. Hence, hierarchical approaches can be found for categorizing images and detecting (complex) objects in images, cf. [1–6], where often instances of general classes, such as, e.g., *airplane*, *building*, *cloth*, *dog*, *face* etc. are to segment and to recognize. A different also very challenging task is the detailed interpretation of terrestrial facade images, i. e., the derivation of a scene description with information about the parts of the recognized building. This task has been attracted by the computer vision community due to the fast developments of virtual 3D city models. So far, such city models with several hundred thousands of buildings are only used for visualization, but the integration of semantics would significantly enrich their purpose. Obviously, the interpretation of facade images should be performed automatically.

Buildings and their parts as windows, doors and balconies are very challenging objects due to their large variety in shape, size, color and texture. To detect such objects in images, many different approaches have been proposed in last years. E.g., main authors (see [7, 8]) try to classify pixels or larger patches using Markov Random Fields (MRF). While their focus lies on separating *building*, *ground*, *sky* and *vegetation* from each other.The contextual scene interpretation by considering different object sizes and therefore image scales for object classification has already been applied in [3, 7, 9], but these approaches either suffer under too simple regions, e.g., patches which cannot be used for describing complex shapes, or they have a very high complexity.

R. Mester and M. Felsberg (Eds.): DAGM 2011, LNCS 6835, pp. 1–10, 2011.

The spatial arrangement of facade elements is also considered in [10, 11] where the authors propose to use spatial grammars for their scene interpretation. In [12–14] the authors propose more ore less successful strategies for recognizing windows in facade images, but rely on the rectangular shape of windows with strong contours and distinctive corners, or they consider the repetitive structure of many windows in building facades. A very simple blob detector has been proposed by [15] who apply a saliency based image analysis. In experiments, they obtained promising detection rates for windows, but other facade parts, especially the smaller ones, have relatively low detection rates.

The success of the window detectors leads to the question, if we could also design reliable detector for other facade parts as roof, doors or balconies. These objects are more challenging due to their more variable appearance in images and their lower frequency. Thus, we are pessimistic that this is a promising strategy. Instead of spending much effort into modeling detectors for such objects, we propose to integrate segmentation results in the scene interpretation. Thereby, we focus on object hierarchies, believing that we obtain better classification results in case we integrate classification results of higher image scales when analyzing lower image scales, e.g., we do not want to look for windows or doors where we believe to see vegetation.

We propose a scene interpretation framework, which can be trained to detect instances of various types. Therefore, we segment image regions at several image scales and arrange them in a hierarchical order. [5] use their hierarchy to derive features from various scales, which are used to build a feature vector for regions of the lowest scale only. In contrast to [5], we also want to classify the regions at higher scales, thus we individually derive features for each segmented region. We improve our classification by an additional analysis using the region hierarchy, which we realize as a conditional Bayesian network. Since Bayesian networks only infer hierarchy information, we also integrate context knowledge by extracting features characterizing the neighborhood of a region. This synthesis of methods enables us to develop a very flexible data-driven scene interpretation approach.

The paper is organized as followed. In sec. 2, we present our concept of a conditional Bayesian network which is constructed by using a hierarchy of segmented regions. Further details to our approach are given in sec. 3. Then, we present our results in sec. 4. In sec. 5, we discuss an extension of our approach for more general scene interpretation tasks. Finally, we summarize our approach and discuss possible extensions of it in sec. 6.

2 Concept for Conditional Bayesian Network

We want to develop a methodology, which is able to detect instances of different classes. These classes may describe well-shaped things, such as *buildings* and their parts as *windows* or *doors*, and formless stuff, such as *sky* or *vegetation*. Due to the facts that we look of objects which can be arranged hierarchically and we are interested in interpreting man-made scenes where we often find precisely detectable object contours, we propose to segment distinctive image regions which are hierarchically ordered to obtain image evidence for further classification. This further step consists of three steps: extracting features for each region, classifying it by a conventional classifier, and finally we construct a conditional Bayesian network to infer information through the hierarchy of regions. At we end, we have consistent classification of image regions, and the classification results can be visualized in the image. Fig. 1 shows at the left side the input image of a building scene in suburban environment, and below of it, the ideal classification results of *building* and *window* are shown. At the right of fig. 1, we show a hierarchy of manually segmented image regions.

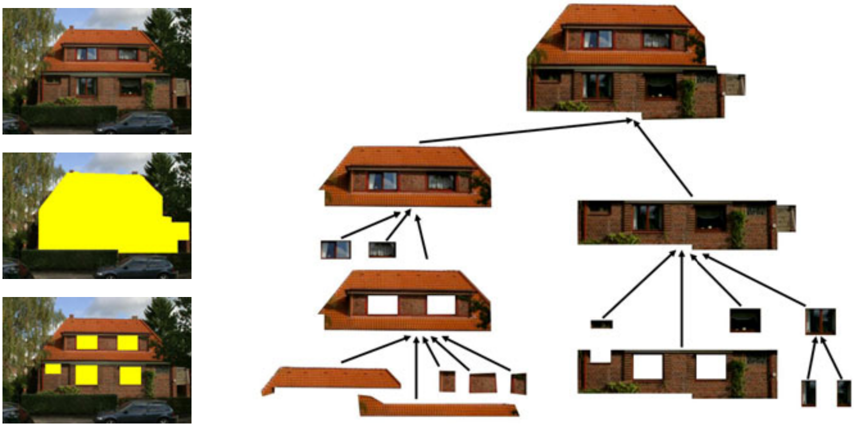


Fig. 1. Left: Facade image (top row) and manually marked objects of class *building* (middle row) and class *window* (bottom row) in yellow. Right: Hierarchical segmentation of *building* object with parts of *roof*, *wall*, and *window*.

We call a segmented image region S_m, and note their hierarchical order by the parent-relation π. I.e. for each region S_m we find exactly one parent $S_{\pi(m)}$, and the parent-relation does not exist for top regions in the hierarchy. Usually, hierarchies are defined by inclusion of smaller elements by larger ones. Hence, the parent of a region $S_{\pi(m)}$ holds information on region S_m and of a distinctive neighborhood. In our point of view, this is more realistic than learning about neighborhoods of all directions as typically done in MRFs.

We derive a block of features F which consists of feature vectors F_m extracted from region S_m. We use the region hierarchy and the block of features F to construct a Bayesian network, as visualized in fig. 2. If we have segmented M regions, the graph of the Bayesian network consists of $M+1$ nodes. For each region, we introduce M random variables $\underline{x}_m$ which are modeled discrete with C states, which describe the probability for the m-th region to belong to one of the C classes. The additional node in the graph is F which is treated like an observed random variable, because the features do not change when inferring in the Bayesian network. Thus, F makes our network to a conditional one.

We obtain the best result for our scene interpretation, if we maximize the probability $P(\underline{x}_1,\ldots,\underline{x}_M,F)$, which we can approximate by

$$
\begin{aligned}
& P(\underline{x}_1,\ldots,\underline{x}_M,F) && (1)\\
&= P(\underline{x}_1,\ldots,\underline{x}_M \mid F)P(F) && (2)\\
&= P(F)P(\underline{x}_1 \mid F)\prod_{m>1} P(\underline{x}_m \mid \underline{x}_{\pi(m)},F) && (3)\\
&= P(F)P(\underline{x}_1 \mid F)\prod_{m>1} P(\underline{x}_m \mid \underline{x}_{\pi(m)})P(\underline{x}_m \mid F) && (4)\\
&\doteq P(F)\prod P(\underline{x}_m \mid F)\prod_{m>1} P(\underline{x}_m \mid \underline{x}_{\pi(m)}) && (5)\\
&\propto \prod P(\underline{x}_m \mid F)\prod_{m>1} P(\underline{x}_m \mid \underline{x}_{\pi(m)}). && (6)
\end{aligned}
$$

The right side of eq. 6 contains only two terms, which we want to derive from training data. Thereby, we approximate $P(\underline{x}_m \mid F)$ by learning a classifier κ on the basis of region-specific features, i.e. by $P(\underline{x}_m \mid F_m)$. Then, classifier κ returns probabilities of region S_m to belong to class c. The other term $P(\underline{x}_m \mid \underline{x}_{\pi(m)})$ can simply be learned from counting class labels of the training data dependent on the region hierarchy.

3 Realization Regarding Facade Image Interpretation

In this section, we describe how we have realized our concept for interpreting facade images. We applied segmentation, feature extraction and classification methods which are either designed with respect to that domain, or they are simple and efficient.

3.1 Hierarchical Segmentation

Several general and domain-specific approaches have recently been proposed to segment facade images. While the authors of [16] proposed a domain-specific strategy with subdividing the scene into rectangles, we developed a more flexible segmentation earlier, cf. [17]. There, we determine watershed regions in a dense scale-space with 41 scales. To reduce the number of regions, we proposed only to select stable regions, i. e. we obtain $M \approx 1000$ stable regions S_m. In experiments on facade images [17], we showed that we are able to detect small objects, such as windows, and larger ones, such as buildings. Furthermore, we showed that the hierarchy of stable regions reflects the object structure.

3.2 Features of Regions

For each region S_m, we extract a D-dimensional feature vector with $D = 65$. We use region-specific features as its area, circumference, form factor, and aspect ratio. Others describe the region and the difference to its neighborhood, e.g. mean and standard deviation of the color channels as well as the color differences. Furthermore, texture features derived from Haar transform, characteristics of the gradients similar to HoG-descriptors by [18], and characteristics of the generalized region by a 4-corner-polygon

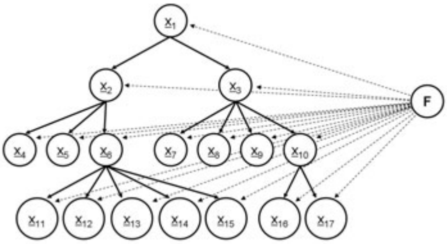


Fig. 2. Conditional Bayesian network derived from hierarchical segmentation of Fig. 1

as its angles or its area ratio to the original region. In preliminary tests, we evaluated that these features are sufficient good for separating different objects. The corresponding class label of each region, the best fitting label $\widehat{x}_m$, is derived from manually labeled annotations, cf. sec. 4.

3.3 Classification of Regions

Here, we describe how we design the classifier κ. Usually, at smaller scales, small regions are segmented, which have homogeneous color or texture, but no characteristic shape. At higher scale, this turns around: shape is often more informative than color or texture. Therefore, we divide our set of segmented regions into subsets. Stability of a region is defined in [17] by only slight changes of a region over several scales. Thus we find each detected region in at least one of our five reference scales in scale space ($\sigma = 1, 2, 4, 8, 16$). The lowest one defines the subset membership for classification.

For each training data subset, we perform a Linear Discriminant Analysis (LDA) which determines the optimal feature subspace for separating the classes. There, we determine class-specific probability density functions (PDF) by mixtures of three Gaussian distributions (GM). Three GM are more reliable than just one, because we must expect very heterogeneous data, e. g., building may have a homogeneously colored wall, and they can be textured by their bricks. Then, we determine for each sample the PDF for each of the C classes, and we obtain the probabilities for the sample's class membership $\underline{x}_m$ after normalization. The class with the highest probability is taken as result of the classification noted by $\widetilde{\underline{x}}_m$.

3.4 Learning Probabilities of Hierarchy

For applying our conditional Bayesian network, we further need to determine probabilities for class appearances in the region hierarchy. Again, we designed the probability by depending it on the scale of region S_m. We derive the probabilities from counting the relationships of targets of hierarchically ordered regions.

The class hierarchy itself might not be sufficient enough. E.g., if you recognize a dark region in a red roof-region, then it is more likely a window than roof tiles, which would be red again. Hence, we decided to specify the probabilities more detailed, and we integrated the features of both regions, S_m and its parent $S_{\pi(m)}$. For each feature, we model the PDF by class-dependent histogram with ten equally filled bins w.r.t. both regions, yielding in a 2D matrix of 100 entries. Evaluating the improvement of the classification by the conditional Bayesian network, we are able to decide which feature to choose for testing our algorithm.

3.5 Conditional Bayesian Network

The segmented regions are arranged in a hierarchy which forms a forest of trees. In the Bayesian network, we model a random variable for each segmented region, which has two parents: the random variable of the parent's region and the random variable describing the region's classification in the LDA-subspace.

Now, we want to determine the best probabilities of all random variables, i.e. the best classification results of all regions. Therefore, we apply the inference algorithm of polytree-structured Bayesian networks as proposed in [19]. Since our structure of hierarchically segmented regions only consists of trees, the inference algorithm is very simple. As result we obtain vectors with C elements, each one reflecting the inferred probabilities for a region's class membership. The class label with the highest probability is noted by $\widetilde{\underline{x}}'_m$ and selected as new classification result.

4 Experiments

4.1 Setup Up of Our Evaluation and Used Data

In the previous sections we described our concept and explained how we have realized it. Now we want to show some results, and evaluate our approach. We tested our classification framework on the benchmark data set by [20], where also regions of objects and the relations of parts are available. The data set contains of 60 facade images and their manually labeled annotations, showing buildings of various sizes and styles, mainly acquired in Germany and Switzerland. We divided the data set into five equally sized subsets with 12 images, which were used for testing, while the other 48 were used for training the classifier and learning the probabilities on hierarchy. By performing a cross validation test, we managed each image being a test image exactly once.

For assigning target values for the regions, i.e., the best fitting class label $\widehat{x}_m$, we first check the manually labeled pixel-wise annotations of [20]. If there is one most dominant class label, we select this label as target for the region. Otherwise, we check, if the region overlaps with different objects, then we call the region *mixture*. Simultaneously, we check, if the regions overlaps with different objects where one is part of the other, then it gets the class label of the superior class. Finally, if the region shows too much image content, which is not labeled in the annotation, we assign this region to be *background*. After determining the target of each region, we merge all classes together, which appear less than 3%, and form a new class *others*. So, we hope to avoid many misclassification due to the low appearance of some classes. In total, we will perform a classification of regions considering the $C = 7$ classes *building*, *window*, *vegetation*, *car*, *mixture*, *background*, *others*. The class *others* contains the regions with the original class labels *door*, *pavement*, *ground* and *sky*.

Table 1. Success and misdetection rates s and d of original classifier κ (LDA with GM) and s' and d' of classification with Bayesian network). p marks the portion of true samples of whole data set.

class	p	s	s'	d	d'
building	0.319	0.488	0.648	0.457	0.4
vegetation	0.245	0.759	0.811	0.427	0.331
window	0.237	0.593	0.729	0.527	0.386
others	0.91	0.215	0.323	0.56	0.392
background	0.40	0.66	0.77	0.833	0.709
car	0.37	0.239	0.227	0.687	0.497
mixture	0.31	0.15	0.002	0.854	0.873

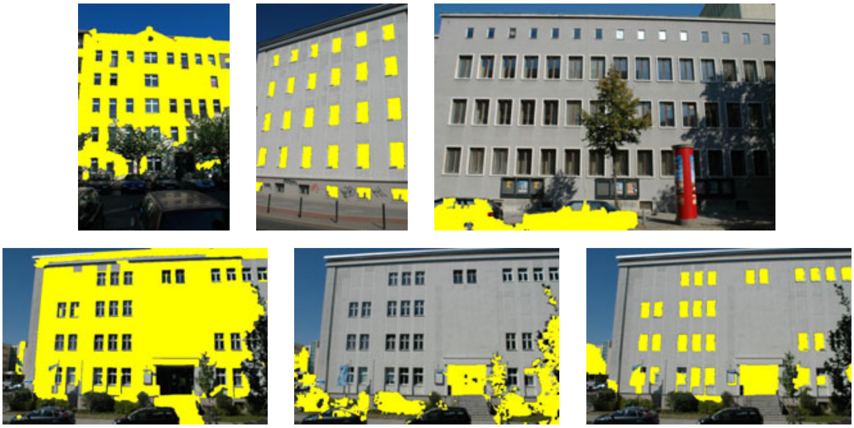
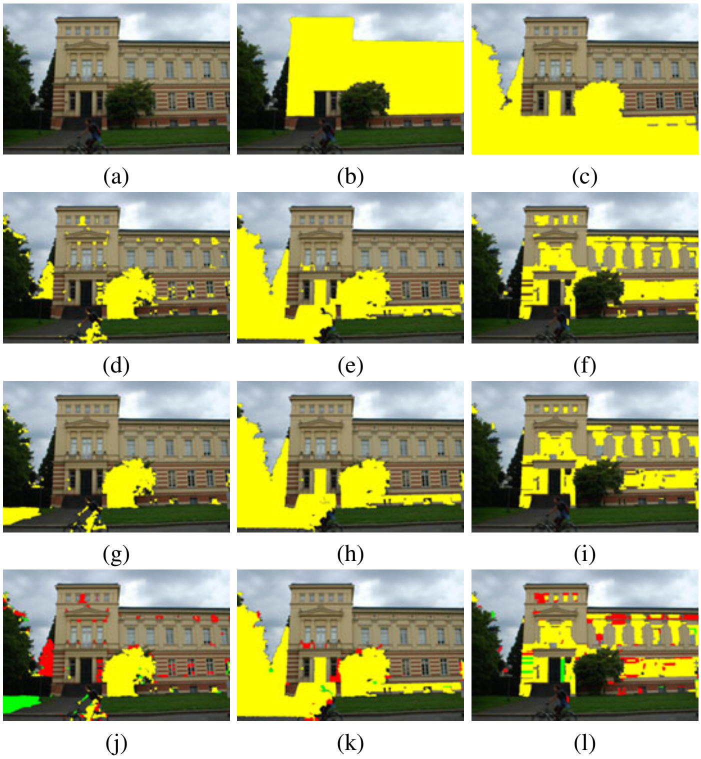


Fig. 3. Four scenes from Berlin (Germany) and classification results from the conditional Bayesian network. Top row: results w.r.t. classes *building*, *window*, and *car*, respectively. Bottom row: w.r.t. classes *building*, *vegetation*, and *window*, respectively.

4.2 Results

In total, our segmentation algorithm segments 131 060 stable regions in 60 images. We tested our approach and obtained the two classification results $\widetilde{x}_m$ and $\widetilde{x}'_m$ for each region and compared them to the the region's target $\widehat{x}_m$. We determined the number of true and false positives, respective, and we define their portion of all true respectively all positively classified samples as success-rate (s or s') and mis-detection rate (d or d'). The prime indicates the classification after inferring the Bayesian network. With classification by classifier κ (LDA and GM) we obtained a success of $s = 0.514$ correctly classified regions, after inferring the information in the Bayesian network, we could improve our success-rate to $s' = 0.620$. The class-specific success-rates of our classification are shown in table 1.

Table 1 also shows the bad classification results for less occurring classes. In our data set, we have 80% of all regions with a label *building*, *vegetation* or *window*. Thus,



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

Fig. 4. Scene in Bonn (Germany). In the top row, (a) shows the original image, (b,c) the output of *building* and *vegetation* at the highest scale, respectively. In the next row, (d,e) show the κ-output for *vegetation* at two different lower scales and (f) the κ-output for *window* at a lower scale. In the bottom row, (g,h) show the output of the Bayesian network for *vegetation* at the same scales as a row above and (i) the output of the Bayesian network for *window* at the same scale as a row above. Last row shows differences between CBN-output and κ-output. Red regions are no longer classified, green regions are newly classified as vegetation or window, respectively.

classifiers typically perform better, if they perform well on these classes. Consequently, low occurring classes as *background*, *car* and *mixture* have very low success rates.

Fig. 3 shows three results from Berlin, Germany, where we obtain really good results with respect to one class. Here we see, that our classification scheme could be used for object detection in images as well. For further visual inspection, we prepared images showing the output of the classifications in fig.4. Within the image part visualized as *building* in (b) the classification results in the lower scales (d,e,f) compared to (g,h,i) improve significantly.

5 Adaptation of the Concept

Our concept for a Bayesian network used for scene interpretation as presented in sec. 2 only relies on (i) ground truth annotations, (ii) a hierarchical segmentation of the scene, (iii) the extraction of features for segmented regions including characteristics on their neighborhood and (iv) a classifier which returns probabilities regarding the region's class membership. Our proposed Conditional Bayesian network remains, although the other components may get changed. So far, we chose these components with respect to the domain of interpreting man-made scenes, where we want to recognize complex objects, such as facades, as well as their parts including their structure.

We are confident that we could transfer our concept to more general scene interpretation tasks as segmenting and classifying objects using the MSRC data set [21] or ImageNet [22]. These data sets do not use overlapping classes, i. e. objects and their parts (building resp. window), but show symbolic image descriptions with a single class for each pixel. The mapping between the segmented regions can be easily adapted, maybe our additional label *mixture* can be dropped. Furthermore, our segmentation could be exchanged by [9, 23], because the analysis of image partitions at a few scales might be more efficient for general recognition tasks than working with selected, but stable regions from various scales. Then, our feature vectors could get extended by additional features, e. g. [18] or [21], and a more powerful classifier as random forests or logistic regression could get integrated. Consequently, the classification results of a reference scale should be selected for its evaluation.

6 Conclusions

We propose a methodology for scene interpretation which combines the hierarchically ordered output of image segmentation and classification on the basis of region-specific features. The tree-structure of the segmented image regions is used to construct a conditional Bayesian network, and we may apply a very efficient inference algorithm. In the conditional Bayesian network, we combine the probabilities reflecting the region's class membership by a common classifier and the class-specific coherences within the hierarchy to improve the classification of segmented regions.

We presented reasonable results for detecting building parts and other objects in terrestrial facade images using the benchmark data set [20] and working with seven classes. We have increased the classification performance of our segmented regions from 51% (ordinary classification) to 62% (classification with conditional Bayesian network). Our approach is very efficient, because we may learn fast the needed probabilities on the hierarchy and the region-specific classification, and the inference of the network is simpler than in common MRFs.

Finally, we discussed how our approach can get generalized for application in other scene interpretation tasks. Further developments w.r.t. the domain of facade images could be done by integrating our results as input of a grammar-based approach. In our point of view, this extension should further improve the results, because they also consider the spatial arrangement of facade elements, e.g., the repetitive structures of windows.

References

1. Epshtein, B., Ullman, S.: Semantic Hierarchies for Recognizing Objects and Parts. In: CVPR (2007)
2. Fidler, S., Leonardis, A.: Towards Scalable Representations of Object Categories: Learning a Hierarchy of Parts. In: CVPR (2007)

3. Schnitzspan, P., Fritz, M., Schiele, B.: Hierarchical support vector random fields: Joint training to combine local and global features. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 527–540. Springer, Heidelberg (2008)
4. Ladický, L., Russell, C., Kohli, P., Torr, P.H.S.: Associative Hierarchical CRFs for Object Class Image Segmentation. In: ICCV, pp. 739–746 (2009)
5. Lim, J.J., Arbeláez, P., Gu, C., Malik, J.: Context by Region Ancestry. In: ICCV (2009)
6. Ommer, B., Buhmann, J.: Learning the Compositional Nature of Visual Object Categories for Recognition. PAMI 32(3), 501–516 (2010)
7. Kumar, S., Hebert, M.: Man-made Structure Detection in Natural Images using a Causal Multiscale Random Field. In: CVPR, vol. I, pp. 119–226 (2003)
8. Verbeek, J., Triggs, B.: Region Classification with Markov Field Aspect Models. In: CVPR (2007)
9. Plath, N., Toussaint, M., Nakajima, S.: Multi-class Image Segmentation using Conditional Random Fields and Global Classification. In: ICML, pp. 817–824 (2009)
10. Dick, A.R., Torr, P.H.S., Cipolla, R.: Modelling and Interpretation of Architecture from Several Images. IJCV 60(2), 111–134 (2004)
11. Ripperda, N., Brenner, C.: Evaluation of Structure Recognition Using Labelled Facade Images. In: Denzler, J., Notni, G., Süße, H. (eds.) Pattern Recognition. LNCS, vol. 5748, pp. 532–541. Springer, Heidelberg (2009)
12. Lee, S.C., Nevatia, R.: Extraction and Integration of Window in a 3D Building Model from Ground View Images. In: CVPR, vol. II, pp. 113–120 (2004)
13. Reznik, S., Mayer, H.: Implicit Shape Models, Self-Diagnosis, and Model Selection for 3D Facade Interpretation. PFG 2008(3), 187–196 (2008)
14. Čech, J., Šára, R.: Languages for Constrained Binary Segmentation based on Maximum Aposteriori Probability Labeling. Intern. J. of Imaging and Technology 19(2), 66–99 (2009)
15. Jahangiri, M., Petrou, M.: Fully Bottom-up Blob Extraction in Building Facades. In: PRIA (2008)
16. Burochin, J.P., Tournaire, O., Paparoditis, N.: An Unsupervised Hierarchical Segmentation of a Facade Building Image in Elementary 2D-Models. In: ISPRS Workshop on Object Extraction for 3D City Models, Road Databases and Traffic Monitoring, pp. 223–228 (2009)
17. Drauschke, M.: An Irregular Pyramid for Multi-scale Analysis of Objects and Their Parts. In: Torsello, A., Escolano, F., Brun, L. (eds.) GbRPR 2009. LNCS, vol. 5534, pp. 293–303. Springer, Heidelberg (2009)
18. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: CVPR, vol. I, pp. 886–893 (2005)
19. Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge (2000)
20. Korč, F., Förstner, W.: eTRIMS Image Database for Interpreting Images of Man-Made Scenes. Technical Report TR-IGG-P-2009-01, IGG University of Bonn (2009)
21. Shotton, J., Winn, J.M., Rother, C., Criminisi, A.: *textonBoost*: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 1–15. Springer, Heidelberg (2006)
22. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR, pp. 248–255 (2009)
23. Arbeláez, P., Maire, M., Fowlkes, C., Malik, J.: Contour Detection and Hierarchical Image Segmentation. PAMI 33(5), 898–916 (2011)