INTERPRETING STRUCTURES IN MAN-MADE SCENES
Combining Low-Level and High-Level Structure Sources

Kasim Terzić, Lothar Hotz
HITeC e.V., Department of Computer Science, University of Hamburg, Hamburg, Germany
terzic@informatik.uni-hamburg.de, hotz@informatik.uni-hamburg.de

Jan Šochman
Center for Machine Perception, Department of Cybernetics Faculty of Elec. Eng.,
Czech Technical University in Prague, Czech Republic
jan.sochman@cmp.felk.cvut.cz

Keywords: Computer vision, scene interpretation, image processing, high-level vision

Abstract: Recognizing structure is an important aspect of interpreting many computer vision domains. Structure can manifest itself both visually, in terms of repeated low-level phenomena, and conceptually, in terms of a high-level compositional hierarchy. In this paper, we demonstrate an approach for combining a low-level repetitive structure detector with a logical high-level interpretation system. We evaluate the performance on a set of images from the building façade domain.

1 INTRODUCTION

The scene interpretation can be improved by exploiting the structural information present in many domains. In this paper we illustrate the use of high-level structural models for combining structured low-level evidence into a complete, coherent interpretation of an image. The paper addresses the task of window detection in complex façade scenes. Windows in a façade are often organised into regular structures but the structure is often complex and there is a lot of variation. There are two sources of structure available: (i) *visual structure*, manifested in repeated visual phenomena in the image, and (ii) *compositional structure*, expressed as part-of relations in the compositional hierarchy of scene objects.

We exploit the visual structure by using a structural component detection algorithm for window detection. It detects *structural components* in the image consisting of horizontally and vertically aligned detections based on automatically detected seeds. The problem with this approach is that it leads to a number of conflicting and overlapping structural components, and many false positives. Each of the the detected structural components explains *some* of the windows, but usually none of them explains *all* the windows.

The compositional structure is modelled in a knowledge base describing the domain of façade scenes. Objects can be parts of aggregates and their arrangement within the aggregates can be modelled manually or learnt allowing more complex relations than simple alignment. However, logical interpretation based on object-level depends on correct detections. Our approach combines the two sources of structure by using a middle-layer component to form a complete interpretation of the façade.

There is a lot of recent interest in the field of computer vision for using high-level context for interpreting scenes from a wide range of domains such as airport activity recognition (Fusier et al., 2007), interpreting building façades (Hotz et al., 2008; Čech and Šára, 2007) or analysing traffic situations (Mohnhaupt and Neumann, 1993; Hummel et al., 2008). However, we are not aware of other current work combining different high-level and low-level sources of structure for interpreting highly structured domains.

Several other recent approaches deal with the integration of low-level image processing and high-level reasoning like e.g. (Zhu and Mumford, 2006) which uses grammar-based models for detecting structure at multiple levels of abstraction including the low level. Structured scenes are also examined in (Seo et al., 2009), where images of parking blocks consisting of parking spots are interpreted. They demonstrate an approach for filling gaps in structures by using interpolation and extrapolation. In contrast to the more domain-specific approach used in that work our approach is based on a declarative representation of the

constraints and object classes, which facilitates the application of the techniques to different domains (see (Hotz and Neumann, 2005) for an example). We show that the combination of high-level and low-level structural models improves the detection of windows in the façade domain.

The paper is structured as follows. First, the visual and conceptual structure models are described in Section 2. The interpretation process used in our experiments is described in Section 3. The results of an evaluation on a set of annotated images are presented in Section 4 and the paper is concluded in Section 5.

2 INTEGRATING DIFFERENT STRUCTURE MODELS

Two sources of structure are used in this paper. The visual structure is exploited by a low-level process to detect repetitive structures in the image. The results of this procedure are referred to as *evidence*. The evidence is mapped into a high-level reasoning system which interprets the scene based on a conceptual model. The instances of high-level models are referred to as *views*. The matching of low-level evidence to high-level views is done by the middle-layer component Matchbox. This section describes the two structure models and outlines our integration strategy.

2.1 Visual Structure Model

The structure component detection algorithm is a greedy procedure for finding sets of self-similar horizontally and vertically aligned windows in a façade image.

The detection process starts by running a sliding window detector over the image. It produces a set of initial window hypotheses – component seeds – $x_i(\alpha,\beta,s)$, $i=1,\ldots,N$, parametrised by their position (α,β) in the image and their scale s. We trained an AdaBoost classifier as the seed detector (Freund and Schapire, 1997). It is intentionally built to be weak (trained on about 100 window images only) to demonstrate the advantage of structural priors for window detection. Consequently, the seed set contains many false positives and usually several windows on the façade are missed.

The structural component growing starts by initialising each component by one seed x_i. When extending a component by a new window, two factors are considered besides the alignment with already included windows: (i) a window model similarity F_W, and (ii) a component elements appearance similarity F_C.

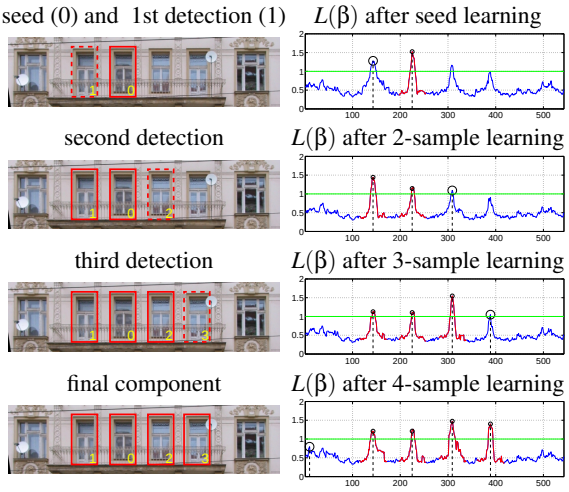


Figure 1: A row component growing process. Right column shows the incrementally updated component confidence L over the row positions β (blue). Already accepted window positions are marked by a small circle in the confidence plot and by a solid-line rectangle in the image (left column). Their neighbourhood which is suppressed for further search is shown as red parts in the confidence plot. A new candidate for the compoment expansion is marked by a big circle in the confidence plot and by a dashed rectangle in the image. The acceptance threshold is plotted in green.

The window model similarity F_W is used to prevent the search to drift away from the window class appearance. In our case, the window class is defined by the AdaBoost classifier. Its real-valued output $f_W(x)=\sum_t h_t(x)$ is used to compute an a posteriori probability estimate of an image x being a window (Friedman et al., 1998)

$$F_W(x)=\frac{\exp(f_W(x))}{\exp(-f_W(x))+\exp(f_W(x))}\,. \qquad (1)$$

The component appearance similarity F_C assures only the windows of the same type are included into a component. It is computed as in equation (1) but with the use of an on-line adapted AdaBoost classifier f_C (Grabner et al., 2006) trained on the compoment windows only. Each time a new window is added to the compoment, the appearance consistency classifier f_C is updated. The negative examples for the update are collected from the window neighbourhood as in (Grabner et al., 2006).

The complete component growing procedure is demonstrated in Figure 1 on a single row of windows. Each component is extended independently in a greedy way maximising a criterion function

$$L(\alpha,\beta,s)=F_W(x(\alpha,\beta,s))+F_C(x(\alpha,\beta,s))\,. \qquad (2)$$

Since the windows in the component are assumed to be aligned and of the same size, only a 1D search

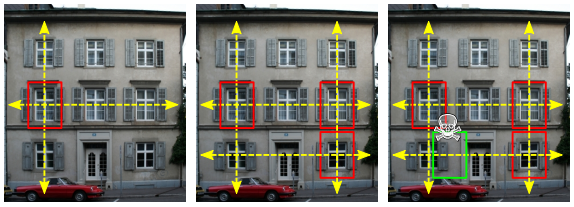


Figure 2: Full 2D growing process with alignment conflict illustration.

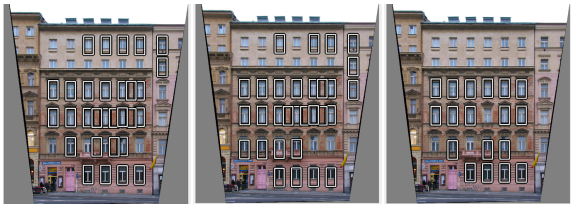


Figure 3: Examples of components detected using sctructure component detection algorithm.

in vertical (α parameter, not used in this example) or horizontal (β parameter) direction is needed. Due to the non-overlapping assumption, the surrounding of already visited positions defined by the window size is masked-out for the search.

To compute the confidence L, the appearance classifier F_C is first initialised by the seed $x_i(\alpha,\beta,s)$. As new windows are added to the component, the classifier is updated incrementally. The search is terminated when $max(L)$ falls below a threshold $\theta = 1$.

Figure 2 shows an example of the component expansion in both vertical and horizontal direction at the same time. For each expansion step a maximum of L is searched along possible horizontal and vertical expansion of all component elements. To preserve the alignment of all component elements, the search is further restricted to positions which are not in conflict with other component windows (Figure 2, right). Several detected components grown from different seeds are shown in Figure 3. The expansion directions used here are particularly suited for façade images, but they can be expressed more generally as a grammar, allowing different types of alignment.

2.2 Compositional Structure Model

As shown in (Hotz and Neumann, 2005), scene interpretation can be formally modelled as a knowledge-based process. Our conceptual model uses a knowledge-based representation of all scene objects that can occur in a façade scene, i.e. primitive objects like `window` or `door`, aggregates like `balcony` or `entrance`, and alignments. Alignments are used

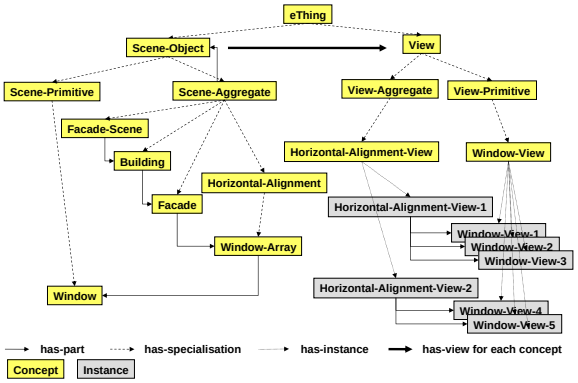


Figure 4: Part of the ontology including instances for an image.

to represent vertical and horizontal structures like a collection of aligned rectangles as they are given by low-level structure detector (see Section 2.1). Such structures can be specialised into alignments like `window-array` or `stack-of-storages`. The approach is general enough to be used on a wide range of domains, and it is applied to windows and window arrays in this case.

Besides the part definitions, the spatial relations (constraints) are given for aggregates and alignments that have to be fulfilled for scene objects of a particular object class. Those constraints are learnt in this case (Hartz et al., 2009) , but they can also be manually constructed. In the case of a window array, the main constraints on the parts are: (i) same size, (ii) same distance, and (iii) vertical alignment, all within a certain tolerance.

The conceptual model represents all known primitive objects, aggregates and alignments, in a declarative way. The interpretation process uses this model for performing interpretation steps that subsequently construct an interpretation of a given scene. The final interpretation of the scene includes the detections provided by the structure as well as hypotheses which are made according to the conceptual model (see Section 3.2).

2.3 The SCENIC System

A realization of a knowledge-based interpretation system capable of interpreting scenes based on models described in the previous section exists in SCENIC (Terzic et al., 2007). SCENIC includes a high-level reasoning layer, a set of low-level image processing modules (IPMs), and a middle-layer component called Matchbox which matches high-level concepts to the detections provided by the IPMs.

The reasoning layer of SCENIC is based on the configuration methodology (Soininen et al., 1998),

which provides commonly known domain independent reasoning techniques like constraint propagation (Yang and Yang, 1997) or rule-based reasoning (Russel and Norvig, 2003) in combination with a declarative domain-specific knowledge base of concepts and constraints. Furthermore, the high-level layer consists of a declarative interpretation process, which structures the reasoning tasks of propagating constraints, instantiating concepts to instances, determining relations between instances, etc. Concepts are mainly aggregate models, their instances represent aggregate instantiations (or simply: "aggregates"), i.e. configurations of concrete objects in scenes. The interpretation process attempts to recognise aggregates, which describe the observed evidence. Figure 4 shows a part of the conceptual model.

Within SCENIC , a strict separation of observed *views* (2D projections within an image) and 3d-objects is given. Views represent typical appearances of scene objects and 3d-objects collect all kinds of information known or inferred for scene objects, e.g. compositional and spatial relations. This allows for occlusion inferences and 3D reasoning, but since this paper focusses on interpreting rectified façade images (façade edges are parallel to image edges), this feature is not exploited in this paper.

The main task of the Matchbox is the grounding of high-level concepts by matching evidence to high-level views. This occurs either in a bottom-up direction, where new views are created from available evidence, or a top-down direction, where view hypotheses are matched to available evidence. In the context of this paper, the bottom-up step is the selection of the best low-level structures and passing them to the high-level. The top-down step is looking for the low-level window detections which fit the hypotheses created by the high-level.

2.4 Integration Approach

The structure component detection algorithm outputs a set of structural components, one for each seed (components containing only a single window are removed). They are overlapping as they represent different parts of the same façade structure grown from different seeds. Since logic-based interpretation systems (in our case SCENIC) expect consistent input, these structures can not all be passed to the high-level at the same time, as they would lead to logical conflicts. Our approach is to integrate the structured evidence by starting with strong evidence and relying on top-down inference to fill in the missing objects. The algorithm is sketched below:

1. select the strongest evidence from the structured bag of evidence by the Matchbox (Section 3.1),
2. interpret the image based on the available evidence (Section 3.2),
3. suggest hypotheses of missing objects (Section 3.2),
4. match the hypotheses to existing unused evidence (Section 3.3).

Steps 3-4 are repeated until all hypotheses are checked and no more hypotheses can be made. The structural information from the low-level. which arranges objects into structures, helps the high-level to instantiate the correct aggregates and thus propagate the necessary constraints. The use of high-level hypotheses means that high-level context is used to pick out the correct evidence from a set of conflicting and spurious detections. This way, both sources of context are combined to create a single interpretation

3 INTERPRETATION PROCESS

3.1 Initialising the High-Level

The low-level evidence consists of many structural components, which in turn consist of many individual window detections. The window detections often overlap with detections from other structural components. We define the confidence C_w of a window detection W_n as

$$C_w(W_n) = \sum_{n \neq m} \frac{Area(W_n \cap W_m)}{max(Area(W_n), Area(W_m))}$$

Based on the confidences of individual window detections W, the confidence C_r is calculated for each window row R as:

$$C_r(R) = \sum_{W_n \in R} C_w(W_n)$$

The rows are then sorted according to their confidence, and a best set of non-overlapping rows is selected as initial evidence for interpretation by using a greedy algorithm. The row with the highest confidence C_r is selected as the best row and all rows overlapping with that row in the vertical dimension are removed. Then the row with the second highest confidence is selected, and the process continues until there are no rows left. Figure 6 shows the result of this algorithm on one example image.

The selected rows and the corresponding window detections are passed to the high level. It is important to note that the partonomical relations are also

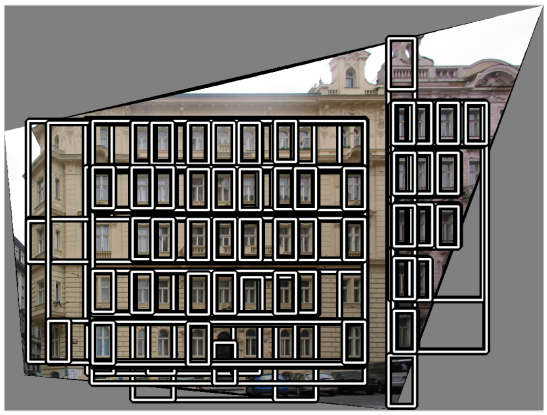


Figure 5: All the evidence from the low level.

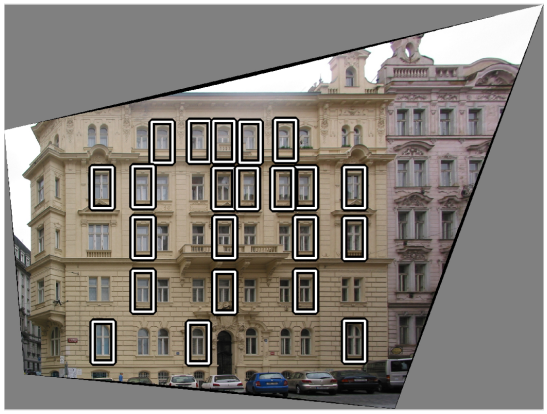


Figure 6: Automatically selected initial evidence. All horizontally aligned windows belong to the same rows (not shown here).

passed as evidence along with the instantiated views, i.e. each window belongs to a row formation along with other windows. This means that the high-level doesn't have to examine all possible combinations of low-level detections in order to find a structure. After the initialization, the high-level interpretation is started.

3.2 High-Level Interpretation Process

The task of the high-level reasoning is to find a logical model for the set of observed scene objects passed as evidence by a lower-level process, i.e. to integrate all scene objects into aggregates corresponding to a conceptual model described in Section 2.2. The interpretation process can hypothesise scene objects if their existence is likely, considering the current interpretation context. All hypotheses made by the interpretation process should be confirmed by the evidence in the scene. For confirmation, a request is sent to the Matchbox, which controls the image processing modules (see Section 3.3 and (Hotz et al., 2008)).

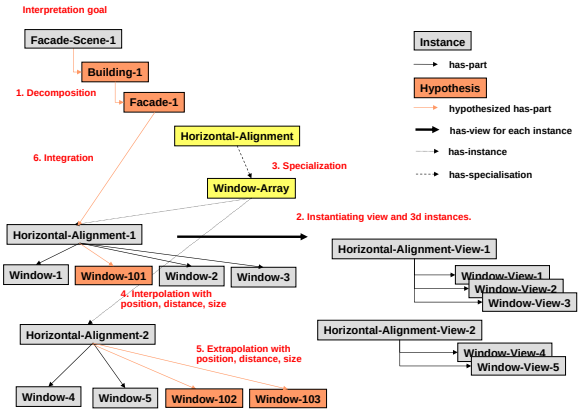


Figure 7: An interpretation process realisation example.

The high-level reasoning used in this paper is realised through a combination of constraint processing, description-logic like inferences, and configuration inferences. See (Hotz and Neumann, 2005) for more details.

The interpretation process starts by creating mandatory parts that directly follow from a given interpretation goal (see Figure 7). In the case of the façade domain, this step is caused by the fact that each façade scene (the goal) has a building which has a façade as a part. These steps are called *decomposition* steps.

As described in the previous section, the next step is to represent structured and primitive views from the Matchbox as instances of view concepts (e.g. `horizontal-alignment-view-1`) and 3d-concepts (e.g. `horizontal-alignment-1`). If the parts of a horizontal alignment are determined by the Matchbox to be of object class `window`, the horizontal alignment can be specialised to a `window-array` (see Step 3). Such window arrays are extended with further window hypotheses based on the the position, size, and distance constraints between the windows of an array. For this step of identifying window hypotheses, three heuristics are applied.

The first one tries to adopt the structure of the nearby window array with the strongest confidence. This structure is considered to be context which might influence positions, distances, and sizes of hypothesised windows. If no context is present, the second heuristic, *interpolation*, fills gaps in a window array by trying to fit window hypotheses between existing windows while observing the window array constraints. The third one, *extrapolation*, hypothesises windows on the left and right side of a window array (see (Seo et al., 2009) for similar steps). As a last step, the observed structure instances are integrated into the façade instance.

The interpretation process creates hypotheses of

missing parts based on the conceptual models and the current scene context. The hypotheses describe the expected objects with their size, position, and object class (in this case, a window). They are matched to available evidence as described in the next section. The hypotheses are updated to reflect the observed evidence position and size ranges and the interpretation process proceeds with those confirmed hypotheses. A result is shown in Figure 8 (c). The white rectangles show the low-level detections selected as initial evidence, and the red rectangles show the hypotheses created by the interpretation process.

3.3 Matching hypotheses to evidence

The hypotheses created as a part of the interpretation process need to be confirmed by evidence from the low level. The task of the Matchbox is to look for the unused evidence which can confirm the generated hypotheses.

The hypotheses are described in terms of allowed ranges for the position and size. The position and the size ranges are computed at the high-level such that the created hypotheses satisfy all the constraints imposed by the high-level window-array model (e.g. vertically aligned windows of similar size, which do non intersect). The hypothesis is confirmed if there is evidence detected in the provided area which has not been passed to the high level during the initial step and which has a size allowed by the size ranges describing the hypothesis. If there are several possible matches, the match with the highest confidence C_w is selected.

Due to the imperfection of low-level detectors, the exact position and size of the detections is extended to ranges of allowed values. These uncertainty ranges cover the observed inaccuracy of the detector and were determined experimentally on a set of annotated images. The matching process then looks at the intersection of the position and size ranges of the hypothesis and the available evidence and confirms the hypothesis if the intersection of all ranges is not empty.

At the end of the interpretation, the unconfirmed hypotheses are discarded as hallucinations, and the combination of the evidence selected in the initial step and the confirmed high-level instances forms the final interpretation of the image.

4 EVALUATION AND DISCUSSION

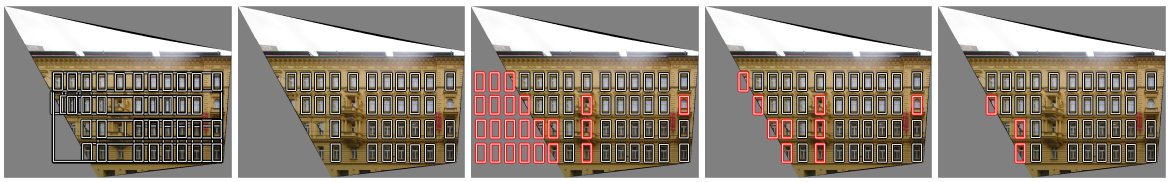
We have tested the combined system on 7 hand-annotated images from the façade domain, consisting of 261 windows.

Table 1 shows the effect of the combined structure models on the detection rate. The selected low-level detections used for initialisation (third column) are tested against the annotation. The detection rate of this bottom-up approach is shown in the sixth column as a baseline for comparison. If the confirmed high-level hypotheses (fifth column) are added, the detection rate generally improves (seventh column).

The different steps of the process can be seen in Figure 8. An interesting observation is that a number of hypotheses which correspond to windows in the image, shown in (d) are not among the hypotheses which were confirmed by the low-level (e). The main cause for this was poor contrast and occlusions which prevented some of the windows from being detected by the low-level stage. In other words, there were no window detections in any of the structural components which corresponded to these windows.

Figure 9 shows the results on the remaining images. In three cases, there was no improvement, since the initially selected evidence has already detected all the windows. The created hypotheses were not confirmed. Image 5 (shown at the bottom) was particularly challenging due to the rich and irregular structure. The confirmed high-level hypotheses detected 15 further windows compared to the bottom-up approach, but also found 3 false positives. The evaluation data show that the combined approach improved the detection considerably on some images (especially images 2, 3 and 5). At the same time, it didn't hurt the cases where the low-level approach was already successful (images 4, 6 and 7). Our combined approach has only resulted in false positive detections in one case (image 5).

False positives can only occur if both structure models expect an object at a wrong location. Since both structure models rely on regularity assumptions, this sometimes occurs on images with an irregular structure. However, as can be seen in Figure 8 (c), the combination of the two structure sources makes good hypotheses that are confirmable by a human. Therefore, we see a potential for further improvement of the detection rate if additional low-level detectors are integrated into the system and used for confirming hypotheses.



(a) (b) (c) (d) (e)

Figure 8: The interpretation process shown on image 3. From left to right: (a) all evidence provided by the structure detector, (b) the evidence used to initialise the high level, (c) all hypotheses (shown in red), (d) the hypotheses which correspond to real objects and (e) the hypotheses which were automatically confirmed by low-level evidence. The windows partially occluded by balconies were hypothesised correctly, but there were no low-level detections to confirm them.

Img1 hypotheses Img2 hypotheses Img4 hypotheses Img6 hypotheses Img7 hypotheses

Img1 confirmed Img2 confirmed Img4 confirmed Img6 confirmed Img7 confirmed

Img5 hypotheses Img5 confirmed

Figure 9: The result of the combined structured models on the testing images.

Table 1: Improvement of the window detection rate after combining two structure sources. In images 4, 6 and 7, the low-level structure detector detected everything except partial windows at the edge of the image, so there was no improvement.

Img	annotated windows	correct low-level windows	correct high-level hypotheses	confirmed correct hypotheses	low-level detection rate	combined detection rate	improvement over pure bottom-up	combined false positives
1	44	27	13	5	61.4%	72.7%	11.3%	0
2	37	23	7	3	62.2%	70.3%	8.1%	0
3	40	33	9	3	82.5%	90%	7.5%	0
4	35	33	4	0	94.3%	94.3%	0%	0
5	60	22	15	15	36.7%	61.7%	25%	3
6	25	24	4	0	96%	96%	0%	0
7	20	19	0	0	95%	95%	0%	0

5 CONCLUSIONS

One of the challenges of scene interpretation is the use of structural information for improving the interpretation of a scene. This paper describes an approach for integrating two separate sources of structure and shows that this combination improves the detection of windows in the façade domain. The low-level structure detector typically computes a large number of potential primitive and structured evidences. A middle-layer component called "Matchbox" reduces this number by selecting the best primitives and structures. High-level reasoning creates hypotheses of missing objects that are caused by the context of the surrounding scene objects. These hypotheses are confirmed or refuted by comparing them to the low-level results. Thus, the Matchbox mediates between both sources of structures, and relates high-level concepts to low-level detections.

The approach was tested on a set of façade images, which are rich in structure. The results show that combining visual and compositional structure can considerably improve the detection of windows in this domain compared to pure bottom-up approach based on visual structure alone. Not all the correct high-level hypotheses were confirmed by low-level evidence, mostly due to poor contrast and partially occluded windows. Further improvements might be possible by using additional low-level detectors for confirming or refuting high-level hypotheses.

ACKNOWLEDGEMENTS

This research has been supported by the European Community under the grant IST 027113, eTRIMS - eTraining for Interpreting Images of Man-Made Scenes.

REFERENCES

Freund, Y. and Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139.

Friedman, J., Hastie, T., and Tibshirani, R. (1998). Additive logistic regression: a statistical view of boosting. Technical report, Department of Statistics, Sequoia Hall, Stanford Univerity.

Fusier, F., Valentin, V., Bremond, F., Thonnat, M., Borg, M., Thirde, D., and Ferryman, J. (2007). Video understanding for complex activity recognition. *Machine Vision and Applications (MVA)*, 18:167–188.

Grabner, H., Grabner, M., and Bischof, H. (2006). Real-time tracking via on-line boosting. In *British Machine Vision Conference*, volume 1, pages 47–56.

Hartz, J., Hotz, L., Neumann, B., and Terzic, K. (2009). Automatic incremental model learning for scene interpretation. In *Proc. of the Fourth IASTED International Conference on Computational Intelligence*, Honolulu, Hawaii.

Hotz, L. and Neumann, B. (2005). Scene Interpretation as a Configuration Task. *Künstliche Intelligenz*, 3:59–65.

Hotz, L., Neumann, B., and Terzic, K. (2008). High-level expectations for low-level image processing. In *KI 2008: Advances in Artificial Intelligence*, volume 5243 of *Springer Lecture Notes in Computer Science*, pages 87–94.

Hummel, B., Thiemann, W., and Lulcheva, I. (2008). Scene understanding of urban road intersections with description logic. In Cohn, A. G., Hogg, D. C., Möller, R., and Neumann, B., editors, *Logic and Probability for Scene Interpretation*, number 08091 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Germany.

Mohnhaupt, M. and Neumann, B. (1993). Understanding object motion: recognition, learning and spatiotemporal reasoning. *Robotics and Autonomous Systems*, pages 65–91.

Russel, S. and Norvig, P. (2003). *Artificial Intelligence - A Modern Approach*. Prentice-Hall.

Seo, Y.-W., Ratliff, N., and Urmson, C. (2009). Self-supervised aerial image analysis for extracting parking lot structure. In *Proc. of Twenty-First Int. Joint Conf. on AI IJCAI-09*, pages 1837–1842, Pasadena.

Soininen, T., Tiihonen, J., Männistö, T., and Sulonen, R. (1998). Towards a General Ontology of Configuration. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing (1998), 12*, pages 357–372.

Terzic, K., Hotz, L., and Neumann, B. (2007). Division of Work During Behaviour Recognition - The SCENIC Approach. In Schuldt, A., editor, *Behaviour Monitoring and Interpretation*, Workshop Proceedings KI, Universität Bremen.

Čech, J. and Šára, R. (2007). Language of the structural models for constrained image segmentation. Technical Report Technical Report TN-eTRIMS-CMP-03-2007, Czech Technical University, Prague.

Yang, C. and Yang, M.-H. (1997). Constraint Networks: A Survey. In *Proc. of the IEEE International Conference on Systems, Man and Cybernetics*, volume 2, Orlando, Florida, USA. Institute of Electrical and Electronics Engineers (IEEE).

Zhu, S. and Mumford, D. (2006). *A Stochastic Grammar of Images*. Foundations and Trends in Computer Graphics and Vision. Prentice-Hall.