Learning to Detect Windows in Urban Environments[1)]

Haider Ali [1], *Christin Seifert* [2], *Nitin Jindal* [2], *Lucas Paletta* [2] *and Gerhard Paar* [2]

[1] *Vienna University of Technology*
Karlsplatz 13, 1040 Wien, Austria
[2] *JOANNEUM RESEARCH - Institute of Digital Image Processing*
Wastiangasse 6, A-8010 Graz, Austria
{*haider.ali,christin.seifert,nitin.jindal,lucas.paletta,gerhard.paar*}*@joanneum.at*

Abstract:
This work is about a novel methodology for window detection in urban environments and its multiple use in vision system applications. The presented method for window detection includes appropriate early image processing, provides a multi-scale Haar wavelet representation for the determination of image tiles which is then fed into a cascaded classifier for the task of window detection. The classifier is learned from a Gentle Adaboost driven cascaded decision tree [1] on masked information from training imagery and is tested towards window based ground truth information which is - together with the original building image databases - publicly available [9, 10, 12]. The experimental results demonstrate that single window detection is to a sufficient degree successful, e.g. for the purpose of building recognition, and, furthermore, that the classifier is in general capable to provide a region of interest operator for the interpretation of urban environments. The extraction of this categorical information is beneficial to index into search spaces for urban object recognition as well as aiming towards providing a semantic focus for accurate post-processing in 3D information processing systems. Targeted applications are (i) mobile services on uncalibrated imagery, e.g. for tourist guidance, (ii) sparse 3D city modeling, and (iii) deformation analysis from high resolution imagery.

1 Introduction

The development of mobile vision systems for the urban context is a challenging research topic today. One major direction of research is towards semantic annotation and indexing into databases from object detection and recognition. A current issue in corresponding activities with respect to computer vision is to focus on the classification of urban infrastructure, such

[1)] This work was funded in part by the EC project MOBVIS (FP6-511051), and the projects "Multi-Sensor Deformation Measurement System Supported by Knowledge Based and Cognitive Vision Techniques" (P18286-N04) and "Cognitive Vision" (S9101) of the Austrian Science Foundation.

as buildings. In this context, the detection and classification of windows can be beneficial for the identification of the respective building, for the processing on the infrastructure geometry, and for the focusing of attention for further interpretation. Other lines of applications are, using the information of window classes for semantic based sparse city modeling, and, applying more accurate processing, e.g., in deformation analysis, on the window based image region of interest.

The original contribution of this paper is to provide a method for window detection and classification in its early stages that could be used for various vision systems and fields of application. We introduce a learning classifier system that provides substantial single window detection and localization, and, at the same time, a window region of interest (WROI) operator that is a basis for further processing. Pattern recognition of windows in urban environments can be mandatory if there is no discriminative texture information available globally that would provide evidence for window occurrences. Building recognition has been proposed in several frameworks [8, 3, 2] and has reached some satisfying level of accuracy. The methods used in their work is mainly relying on the extraction of local information, such as, orientation histograms and SIFT descriptors [3], attention based recognition on selected descriptors [6], or selection of informative descriptors [2]. [8] used the method to identify rectangular features from projected outliers on a rectified wall-plane for window detection, still requiring multiple images for plane estimation. However, to the best of the knowledge of the authors, pattern recognition for rapid window detection has not been considered so far.

The main application for the window detection system will be building, or facade classification from mobile imagery. For this purpose it suffices to detect only a fraction of all windows on a facade, assuming that either the complete set of windows would belong to a single window class, or that the detected windows are sufficiently discriminative with respect to the identification of the corresponding facade so that the remaining undetected windows provide only redundant information. In experiments in view of the targeted applications, the best results achieved so far in the first fair performance evaluation on reference datasets – gained detection rates about 66% of detected windows with respect to ground truth information, and over the complete image data set. However, for building classification we would *need only a fraction of all windows to be detected*. In this context, the proposed window detector would be highly useful by providing sufficient information for further processing. An outlook on further applications that could take advantage of window detection is outlined in Sec. 4.

Window detection as described in this paper is a first step towards a multimodal information based detection system, including the consideration of geometric configurations of collinear lines, exploiting geometric regularity in urban facades, etc. We later intend to integrate evidences from line groupings, pattern detecion and gradient settings into most probable hypotheses on window locations. Window detection as outlined in this paper is currently

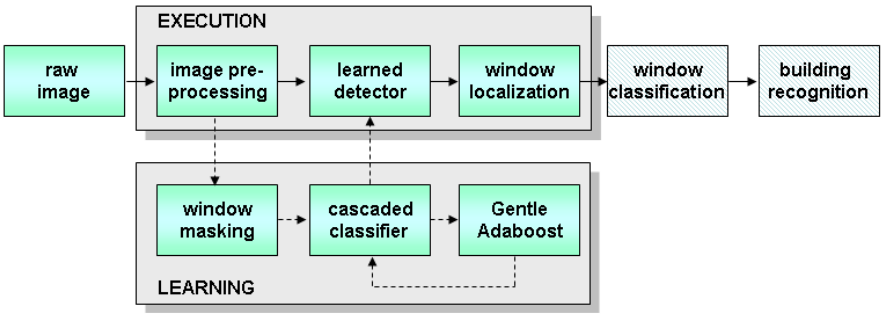


Figure 1: Schematic outline of the window detection system. Dashed lines refer to the learning system. Hatched boxes refer to on-going, not yet reported work.

available as desktop solution but it is planned to integrate it into the client-server solution for building recognition described in detail in [2].

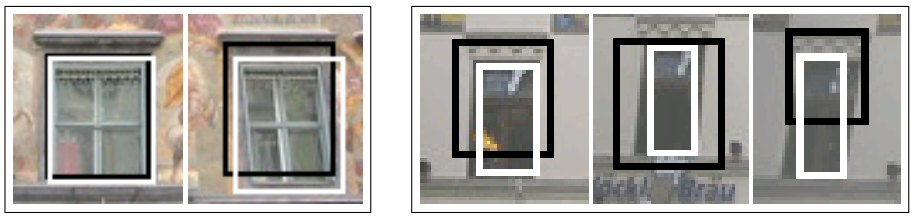
2 Window Detection System

2.1 Overview

The window detection system (Fig. 1) is outlined in a pipeline for training and testing related processing components. Once the cascaded classifier has been learned applying the Adaboost method, it is directly applied to the pre-processed image data. The output of the execution module is a list of coordinates of the bounding boxes of hypothesized window related subimages with respect to the original image frame.

2.2 Learning Window Detection in a Cascade of Classifiers

In this work we propose to formulate the issue of window detection as a pattern recognition problem. In general, it is obvious that the visual content of window patterns might represent considerable variations in their appearance, and that both projective distortion and scale variance due to variation of the viewpoint can have a significant impact on the quality of pattern matching. However, for applications, such as, mobile vision systems, rapid processing and provision of seminal hypotheses it is a relevant issue. Therefore, in the presented work, our interest is on the potential of raw pattern detectors with respect to their capability to provide approximately correct regions of interest for further processing. Consequently, any post-processing, such as, improved matching under consideration of affine or projective transformations might be considered for further optimization in the future. As choice for the learning methodology that is applied for the purpose of window detection we decided to use the work presented by Viola & Jones [11] for detection of objects of interest, under specific consideration of its extention by Lienhart et. al. [4]. The referred object detection system uses Haar-like features and their respective rotated versions. A feature f



(a) (b)

Figure 2: The evaluation of the window detector was based on the quantification of the overlap between the actual window (ground truth, red bounding box) and the localization that is hypothesized by the window detector (green bounding box). Sample cases of *positive true* evaluation for (a) single window (SW) based and (b) windows region of interest (WROI) based detection.

is calculated using the intensity information within two rectangles. A rectangle r is defined by its position (x, y), its width w, its height h and the rotation angle α. For two rectangles $r_1 = (x_1, y_1, w_1, h_1, \alpha_1)$ and $r_2 = (x_2, y_2, w_2, h_2, \alpha_2)$, $\quad 0 \leq x_1, x_2, x_1 + w_1, x_2 + w_2 \leq W$ and $0 \leq y_1, y_2, y_1 + h_1, y_2 + h_2 \leq H$, with W and H are the width and the height of the image, the feature f is defined as $f = w_1 * sum(r_1) + w_2 * sum(r_2)$. $sum(r_i)$ denotes the sum of all pixel intensities within rectangle r_i. The set of all possible features is further reduced by taking only rectangle combinations that mimic early features in the human visual pathway, e.g. edge features. To give an example, for a window of size $24x24$, there are in total 117,941 possible features. These features can be efficiently computed (in constant time) by using the auxiliary images summed area table image (SAT) and rotated summed area table image (RSAT). The binary detection classifier is trained using Gentle Adaboost [1] as the supervised learning methodology. Boosting combines many weak classifiers to one strong classifier. The weak classifiers are only required to be better than chance. The input to the learning algorithm is a set of feature vectors f_i combined with a target class label t_i for each feature vector. $t_i = 1$ if x_i belongs to a complete class of interest (a window) and $t_i = 0$ otherwise. Each weak classifier is trained to reject a certain fraction of non-object patterns. Boosting selects one weak classifier per round that best classifies the weighted training set. After each round of boosting the training set is re-weighted to give the mis-classified samples a higher impact for the next boosting round. During detection a sliding window is moved over the test image across all scales. Each weak classifier defines which feature it is attending on, i.e. the height, the width and the orientation of the two rectangles and their relative position. If a weak classifier accepts a feature, the subwindow is passed to the next stage and the next feature is calculated. If a weak classifier rejects a feature, the subwindow is discarded from further processing. Subwindows that passed the whole classifier cascade are finally classified as containing an object of interest (i.e., a window).

2.3 A Framework for Evaluation

We applied two different evaluation methods, in particular, (i) single window (SW) evaluation and (ii) window region of interest (WROI) evaluation. For SW evaluation, we decided to count a positive true sample if one detection rectangle would be found inside a mask rectangle or, at the maximum, would overlap it by only a few image pixels in each direction. For WROI evaluation, a detection should be counted positive true whenever the mask rectangle would be covered at the minimum by $cov\%$ pixels, e.g., $cov \in \{15, 50, 75\}$. See Fig. 2 for a visualization of the positive true definitions, in both cases. False positives were defined to occur if a mask rectangle is covered by less than 5% of detection pixels in both cases.

3 Experiments and Results

3.1 Training Stage

For the training of the classifier we used international benchmarking image databases that are referred to in the literature regarding building recognition [9, 10, 12]. We manually cut out 1506 windows from the individual databases ZuBuD, TSG-20, and TSG-60. The ZuBuD database consists of images of 201 buildings from Zurich, for each image 5 different views are provided. In the TSG-20 database images of 20 different buildings from the city center of Graz are included. For the training stage we used the images taken from a viewpoint change of $\approx 30^\circ$ compared to the frontal view. We added windows from 40 more images taken from buildings in Graz (part of the TSG-60 database) and from 128 images captured in Vienna. For each window we added the vertically flipped version. Altogether the classifier was trained with 3012 positive (window-) samples and 2524 negative (arbitrary non-window-) samples.

3.2 Evaluation Results

For the evaluation of the window detection system we again used images from the TSG-20, TSG-60 and ZuBuD database (Section 3.1). These images were taken from other viewpoints and under other illumination conditions than the images used for training the classifier, and all downsampled to a resolution of 320x240. We created a ground truth by manually masking all windows in all test images with bounding rectangles. The ground truth then consisted of 744 windows from 40 images of the TSG-20 database, 730 windows from 40 images of the TSG-60 database and 3074 windows from the 115 query images of the ZuBuD database. We applied the two evaluation methods, i.e., SW-evaluation and WROI-evaluation as described in Section 2.3. For SW evaluation a positive true was counted, if one detection rectangle is inside a mask rectangle or at the maximum is overlapping it by 5 pixels in each direction and covers at least 75% of the mask. For WROI evaluation the mask rectangle was covered by 75% by

(a) (b) (c)

(d) (e) (f)

Figure 3: Window detection results. (a) Masked ground truth and (d) good detection results on a TSG-20 [9] test image sample. (b) Masked ground truth and (e) detection on a TSG-60 [10] sample. Decrease in the performance was experienced with (c) modern type windows and buildings with few informative features, and with (f) projective distortions in the window pattern.

Table 1: SW/WROI-Evaluation for TSG-20, TSG-60 and ZuBuD database

Database	PT in 75% Coverage	FP in 75% Coverage
SW TSG-20	57	7
SW TSG-60	52	8
SW ZuBud	30	2
WROI TSG-20	60	7
WROI TSG-60	56	8
WROI Zubud	32	2

a detection to be counted as positive true. Figure 3 shows example test images overlaid with ground truth of window pattern outlines, and the corresponding results for efficient detection.

Single Window (SW) based Evaluation First, we evaluated performance of the window detection system on the 3 different databases using the single window evaluation method (see Sec. 2.3). On the TSG-20 database we achieved the best results, obtaining a positive true accuracy of 57% for a coverage of 75% (Table 1, Section 2.3). On the TSG-60 images

the results are slightly worse, the recognition rate here is 52% and on the ZuBuD database only 30% of all windows were detected. In all cases, we observed that a (huge) variation of the requested coverage had almost no influence on the positive true rate. This means, if a window was detected, at the same time it nicely covered the mask rectangle area. On all three databases there were less than 10% false detections, 7% for the TSG-20, 8% for the TSG-60 and 2% for ZuBuD.

Window Region of Interest (WROI) based Evaluation Since the definition for positive trues is less strict for the WROI evaluation case, we achieved higher detection rates. These vary from 60% for TSG-20 and 56% for TSG-60 to 32% for the ZuBuD database with $cov = 75\%$.

Discussion The differences in the recognition rate in both evaluation cases on the three databases may be explained as follows: First, the TSG-20 contains mostly frontal views of the building, whereas the ZuBuD shows buildings from more extreme viewpoint angles, resulting in affine distortions and large scale variances for single windows (e.g., see Fig. 3). TSG-60 and ZuBuD also contain more modern-type, less structured windows. We argue that, using the window detection system for recognizing buildings from window classes, it is not required to detect all windows of a building but instead to rely on a sufficient number of detections. From the statistical variation of the detection rate per image we understand that we actually receive sufficient window detections: e.g., for TSG-20 the detection rate per image is $57 \pm 19\%$. Similar results for TSG-60 and ZuBuD applied, indicating sufficient support for window classification in potential post-processing.

4 Conclusions

The main application for the window detection system will be building, or facade classification from mobile imagery. For this purpose it suffices to detect only a fraction of windows on a facade, assuming that the complete set of windows would belong to a single window class, or that the detected windows are sufficiently discriminative with respect to the identification of the corresponding facade so that undetected windows would provide only redundant information. The presented detection system actually proved to be capable of providing substantial contextual information for building recognition. Another track of future application will be to apply the detection of windows on facades as a semantic information based preprocessing tool for city reconstruction, such as [5]. Similar systems contribute in the todays growing field of deformation analysis in areas with old buildings like in Europe and East Asia [7]. We presented a framework for learning to detect windows from mobile imagery in the urban environment, based on Gentle Adaboost to optimize a cascaded classifier for detection. The argument for selecting a pattern recognition based methodology for detection was to enable

rapid indexing into the visual information in the context of mobile vision systems, and to provide an interest operator for applications that would post-process the ROI for accurate 3D information processing. The experimental results on standard benchmarking image databases are satisfactory, fulfilling the requirement that at least few windows per building should be detected to enable window classification for context driven building recognition. However, it is noted that the system provides better results on more textured information, such as, old-type windows, and on frontal views of windows.

References

[1] Yoav Freund and Robert E. Schapire. Experiments with a new boosting algorithm. In *International Conference on Machine Learning*, pages 148–156, 1996.

[2] Gerald Fritz, Christin Seifert, and Lucas Paletta. A Mobile Vision System for Urban Object Detection with Informative Local Descriptors. In *Proc. IEEE 4th International Conference on Computer Vision Systems, ICVS*, New York, NY, January 2006.

[3] Jana Kosecka and Wei Zhang. Extraction, matching, and pose recovery based on dominant rectangular structures. *Computer Vision and Image Understanding*, 100(3):274–293, December 2005.

[4] Rainer Lienhart, Alexander Kuranov, and Vadim Pisarevsky. Empirical Analysis of Detection Cascades of Boosted Classifiers for Rapid Object Detection. Technical report, Microprocessor Research Lab, Intel Labs, Intel Corporation, Santa Clara, CA 95052, USA, May 2002.

[5] P. Mueller, P. Wonka, S. Haegler, A. Ulmer, and L. Van Gool. Procedural modeling of buildings. In *Proceedings of ACM SIGGRAPH 2006 / ACM Transactions on Graphics (TOG)*, volume 25, pages 614–623. ACM Press, 2006.

[6] Lucas Paletta, Gerald Fritz, and Christin Seifert. Q-Learning of Sequential Attention for Visual Object Recognition from Informative Local Descriptors. In *Proc. 22nd International Conference on Machine Learning, ICML 2005*, pages 649–656, Bonn, Germany, August 7–11 2005.

[7] A. Reiterer. A semi-automatic image-based measurement system. In *Proceedings of Image Engineering and Vision Metrology*, Dredsden, Germany, 2006.

[8] Konrad Schindler and Joachim Bauer. A model-based method for building reconstruction. In *HLK '03: Proceedings of the First IEEE International Workshop on Higher-Level Knowledge in 3D Modeling and Motion Analysis*, page 74, Washington, DC, USA, 2003. IEEE Computer Society.

[9] TSG-20: Tourist Sights Graz Image Database. http://dib.joanneum.at/cape/TSG-20/.

[10] TSG-60: Tourist Sights Graz Image Database. http://dib.joanneum.at/cape/TSG-60/.

[11] Paul Viola and Michael Jones. Rapid Object Detection using a Boosted Cascade of Simple Features. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001.

[12] ZuBuD: Zurich Building Image Database. http://www.vision.ee.ethz.ch/showroom/zubud/index.en.html.