AUTOMATED APPEARANCE-BASED BUILDING DETECTION IN TERRESTRIAL IMAGES

Jan Böhm*, Norbert Haala, Peter Kapusy

Institute for Photogrammetry (ifp), University of Stuttgart, Germany
Geschwister-Scholl-Strasse 24D, D-70174 Stuttgart
Jan.Boehm@ifp.uni-stuttgart.de

KEY WORDS: Object Recognition, CAD, Orientation, Navigation, Augmented Reality

ABSTRACT:

We present a method for automated appearance-based detection of buildings in terrestrial images. The problem is stated as follows: From an image with a given approximated exterior orientation and a three-dimensional CAD Model of the building, detect the exact location of the building in the image. The method we have developed uses the combination of an imaging device and hardware to approximately measure the exterior orientation. This paper presents a combination of our work on close-range photogrammetry and virtual city models.

1. INTRODUCTION

With the increased availability of low-cost and low-power consumption imaging devices we see a strong increase in their integration into mobile devices such as laptops, personal digital assistants (PDAs), mobile phones and so on. The combination of mobile computational capabilities, imaging capabilities, positioning capabilities and network access opens the door for a variety of novel applications, such as pedestrian navigation aids, mobile information systems and others usually referred to as 'location-based services'. With these devices at hand our interest is to exploit the capabilities of the imaging device for photogrammetric processing.

Personal navigation and the provision of location dependent information are key features of location aware applications. One option to reach this goal is the application of augmented reality (AR) techniques. These techniques are based on the overlay of computer-generated graphics to the user's actual view. The computer graphics are generated based on a spatial model of the visible environment. Of course the virtual computer graphic objects have to be overlaid to their corresponding primitives in the real world as the user observes them. For this reason the accurate determination of the actual position and orientation of the user is required in order to enable a precise mapping of the data. This paper presents a method for the precise detection and localization of buildings in terrestrial images and thereby also provides the means for better navigation of users.

Within an urban environment AR can for example be applied for the presentation of name labels or additional alphanumeric data appearing to be attached to a side of a building. In addition to the visualization of these virtual signposts, more specialized applications could aim at the display of information based on "X-ray vision" in order to present features normally not visible for the user. Typical objects of interest are features hidden behind the facades of a building like the location of rooms or information on infrastructure like the position of power-lines. The integration of augmented reality into a tourist information system is another application for this kind of technique. The so-called telepointing capability is an important feature in implementing a intuitive user interface for location based services (Fritsch et al. 2000).

In order to detect objects usually a model of the object has to be available. In our case we need a model of the buildings to be detected. Therefore one of the key components of our approach is a 3D city model. The development of tools for the efficient collection of 3D city models has been a topic of intense research for the past years. In addition to Digital Height Models and data representing streets and urban vegetation, building models are the most important part thereof. Meanwhile a number of algorithms based on 3D measurement from aerial stereo imagery or airborne laser scanner data are available for automatic and semi-automatic collection of 3D building models. A good overview on the current state-of-the-art of experimental systems and commercial software packages is given in (Baltsavias et al. 2001).

Within this article we present a method to analyze the image of an urban scene and reliably and robustly detect buildings. The method uses the combination of an imaging device and hardware to approximately measure the exterior orientation. The result of the detection can be used both for AR applications and for navigation purposes. Section 2 discusses the approach to the problem of object recognition we have chosen. It contains the description of the model representation, the feature extraction and the recognition algorithm. The utilization of the object recognition for the determination of the orientation parameters using spatial resection is detailed in section 3. In Section 4 we present the results of some typical test cases and demonstrate the accuracy of our approach.

2. OBJECT RECOGNITION

When the task is to detect a three-dimensional shape in an image, two general strategies for object representation are available. One is the three dimensional representation of the object which leads to a 3D to 2D matching problem, the other is a two-dimensional representation which leads to a 2D to 2D matching problem. While the former is the more general and theoretically more appealing approach, there are several

* Corresponding author

practical problems, which often prevent its use. One of the problems is the reliability of feature extraction, the other the exponential complexity of the matching task. For the later approach in order to have a two-dimensional representation of a three-dimensional shape, one has to decompose the shape into several views and store a two-dimensional representation for each view. This is referred to as an aspect-graph. For our system, since we have an approximated exterior orientation of the imaging device, we do not have to build the whole aspect graph, rather we create on-the-fly a single view of the shape corresponding to the orientation data, as we receive the image.

2.1 View Class Representation

With the exception of a few simple shapes (a sphere for example) every three dimensional object has a different appearance when seen from a different viewpoint. The more complex the shape is the stronger are the differences. This is the case for buildings. In the view class representation scheme (Koenderink & van Doorn 1979) the space of possible viewpoints is partitioned into view classes. The view classes are arranged in a graph known as the aspect graph (see Figure 1). Each node represents a single view class, each arc represents the transition from one viewpoint to another. The methods used to compute the view classes of an object can be very complex and the aspect graph of a non-trivial shape is quite large. In the subsequent matching process the input data has to be compared to a multitude of nodes.

In our framework we make use of the knowledge of an approximated exterior orientation. This translates to a single point in the space of possible viewpoints. Therefore we are not required to compute the full aspect graph. If the viewpoint is known only to a very low accuracy it is sufficient to compute a small part of the aspect graph in the neighborhood of the approximated viewpoint. Our experience shows that for our application it is usually sufficient to reduce the graph to a single node. This single node does not need to be stored beforehand but can be computed on-the-fly. The matching process is restricted to only one instance further reducing computational costs.

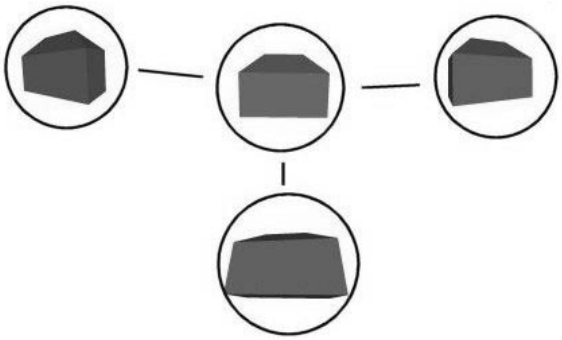

Figure 1: Part of the aspect graph of a simple building.

2.2 Feature Extraction

When designing an object recognition system one has to choose the type of features used for recognition and the algorithm used to perform the matching. The decision on the feature type is often guided by the available model data. In our case the buildings are modeled as polyhedrons, no in-plane facade detail or texture information is available. The strong discrepancy in feature detail in-between model and sensor data prevented us from using edge or corner detection. Since there is no texture information available, image correlation was also not an option.

To achieve a robust detection we chose to detect the overall shape of the building in the image rather than extracting single features. The intent was, that the overall shape is more robust against clutter of the scene, partial occlusion by trees, cars, pedestrians and other negative influences onto the scene. The silhouette of a building is a good representation for its overall shape. From an existing CAD database the CAD model of the building is rendered for a given view according to the calibration data of the camera. The details on how to select the specific building are given in (Klinec & Fritsch 2001). The 'virtual view' of the building is used to extract the silhouette of the building (see Figure 2). This representation is then detected in the scene.

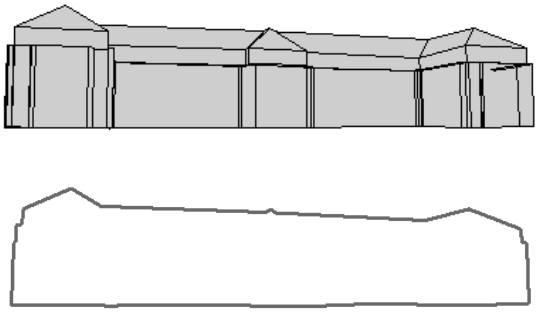

Figure 2: CAD model of a building rendered for a given exterior orientation and the extracted silhouette below.

2.3 Generalized Hough Transform

We decided to use Generalized Hough Transform (GHT) to implement the detection. The GHT is a framework for both the representation and detection of two-dimensional shapes in images. Using the GHT we are able to detect the shape no matter whether it is shifted, rotated or optionally even scaled in the image. We need these degrees of freedom, since the orientation is only known approximately. Additionally the GHT allows for a certain tolerance in shape deviation, which is necessary, since the CAD model of the building is only a coarse generalization of its actual shape as it appears in the image.

The well known conventional Hough Transform (Hough, 1962) is used to detect analytical curves such as lines or circles in an image. The requirement for the application of the Hough Transform is that the model can be formulated as a function of a set of parameters. The GHT is the generalization of that concept for the detection of arbitrary shapes for which a simple analytic description is not possible (Ballard, 1981).

The GHT uses a gradient operator to compute edge magnitude and gradient direction. During the offline phase, which has to be computed once for a certain view class, the so-called R-table is built. First a reference point of the shape is selected usually the centroid of the shape is used. Then the distance vector **r** of every edge pixel **P** to the reference point **O** with respect to its gradient direction **Φ** (see Figure 3) is stored in the R-table. Of course when iterating over all edge pixels, there can be several **r** vectors for an entry of **Φ.**

In the online phase, when the actual detection is performed, all gradients of the search images are computed. For each edge pixel the gradient direction points to an entry in the R-table,

which possibly holds several, vectors $\mathbf{r_j}$. These vectors correspond to positions in the image, which receive a vote. The position in the image receiving the most votes at the end is selected as the position of the shape in the image.
If the orientation of the object is allowed to vary, as is the case in our application, a separate R-table has to be computed for each discrete rotation angle. The same is true for scaling. Thus the formation of the R-tables in our case is quite complex and computationally expensive.

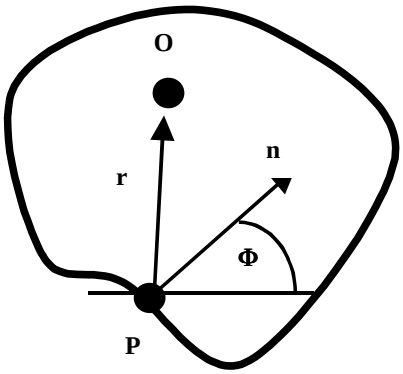

Figure 3: Properties of a single edge pixel P as recorded in the framework of the Generalized Hough Transform.

For our implementation we made use of the HALCON image processing environment, which provides a shape detection mechanism based on the GHT (Ulrich, et. al. 2001). In order to compensate for the computational costs of large R-tables, this operator includes several modifications to the original GHT. For example it uses a hierarchical strategy generating image pyramids to reduce the size of the tables. By transferring approximation values to the next pyramid level the search space is drastically reduced.

3. ORIENTATION

The result of the GHT is a two-dimensional similarity transformation consisting of translation, rotation and scale. This transformation corrects for the misalignment of the shape to its actual appearance in the image. In other words for each three-dimensional point of the original CAD model rendered to the two dimensional image we can use the transformation to correct its location in the image.

3.1 Spatial Resection

This gives us the possibility to create point wise pseudo observations. We can select an arbitrary point from the CAD model of the building and project it onto the image using the calibration data of the camera and the initial approximated orientation. Then we use the two-dimensional transformation from the GHT to move the image point to its correct location. From a minimal configuration of three points we can compute an improved exterior orientation of the camera by photogrammetric spatial resection.
Several alternatives for a closed form solution to the resection problem are given in literature. We follow the approach suggested by (Fischler & Bolles 1981). Named the “Perspective 4 Point Problem” their algorithm solves for the three unknown coordinates of the projection center when the coordinates of four control points lying in a common plane are given. Because the control points are all located on a common plane the mapping in-between image- and object points is a simple plane-to-plane transformation T. The location of the projection center can be extracted from this transformation T when the principal distance of the camera is known. For a detailed description of the formulas please refer to the original publication.
To complete the solution of the resection problem we also need the orientation of the camera in addition to its location. (Kraus 1996) gives the solution for determining the orientation angles when the coordinates of the projection center are already known. The algorithm makes use of only three of the four points.

In principle, the complete process, extraction of building silhouette, improvement of image coordinates by GHT and spatial resection can be repeated iteratively in order to avoid errors resulting from the initial orientation data. Nevertheless, for our application the differences between the projected wire-frame and the image were mainly caused by errors within the available model due to measurement errors or generalization effects. Subsequent iterations did not enhance the result significantly.

4. EXPERIMENTS AND RESULTS

A series of real-world examples was chosen from our database, to investigate both the feasibility of our approach and the accuracy obtained from photogrammetric processing. Our investigations are based on a dataset of the city of Stuttgart provided by the City Surveying Office of Stuttgart. This data was collected by manual photogrammetric stereo measurement from images at 1:10000 scale (Wolf 1999). For data collection the public Automated Real Estate Map (ALK) was additionally used. Thus, a horizontal accuracy in the centimeter level as well as a large amount of detail is available.

Figure 4: Prototype of the mobile photogrammetry device consisting of a camera, a compass and a tilt sensor. The GPS is mounted separately. The laser unit was not used in this project.

4.1 Hardware

The platform we used in order to carry out our experiments is a prototype device for mobile photogrammetry (see Figure 4). It consists of a standard resolution color video camera with extreme wide-angle lens, a GPS receiver, an electronic compass and a tilt sensor. By combining image data and orientation data we obtain an image with an approximate exterior orientation.
The camera is a consumer style Sony DFW-500 video camera connected to the system via IEEE 1394 also known as FireWire. The camera was calibrated on a test-field using a ten parameter camera model, bundle adjustment was performed using Australis (Fraser, C. S. 1997). The additional parameters obtained were used for the computation of the resection, while rendering was done with a reduced parameter set. The GPS receiver is a Garmin LP-25, which can be operated both in normal and differential mode. We used the ALF service (Accurate Positioning by Low Frequency) of Deutsche Telekom for differential mode, thus obtaining a correction signal every three seconds. Our experience shows that the system allows for a determination of the exterior orientation of the camera to a precision of 7-10 m in planar coordinates. The orientation accuracy provided by the digital compass and the tilt sensor resulted in an error of 1° – 2°.

All the devices are connected to a laptop. While the camera and compass/tilt sensor are hand held, the GPS is attached to a backpack. Both the model computation and the image processing was implemented on a standard PC / laptop.

4.2 Recognition Results

Two examples for the recognition of buildings using our approach are given with the opera house in Figure 5 and with a museum in Figure 6. In both examples perfect matches were obtained. This is the case even though one can clearly see the problems, which are encountered in realistic situations. For the opera house a tree is obstructing the view partly occluding the building. A shadow is cast across the facade generating additional gradients. The statues on the roof are not represented in the model. Still the algorithm is able to detect the overall shape of the building, demonstrating the robustness of our approach.
While the initial position of the silhouette in the image in Figure 5 seems correct, it appears to be too large indicating an approximated position in object space too close to the building. For Figure 6 we can see a clear misalignment of the silhouette caused by false orientation. The method is able to correct these errors in both cases.

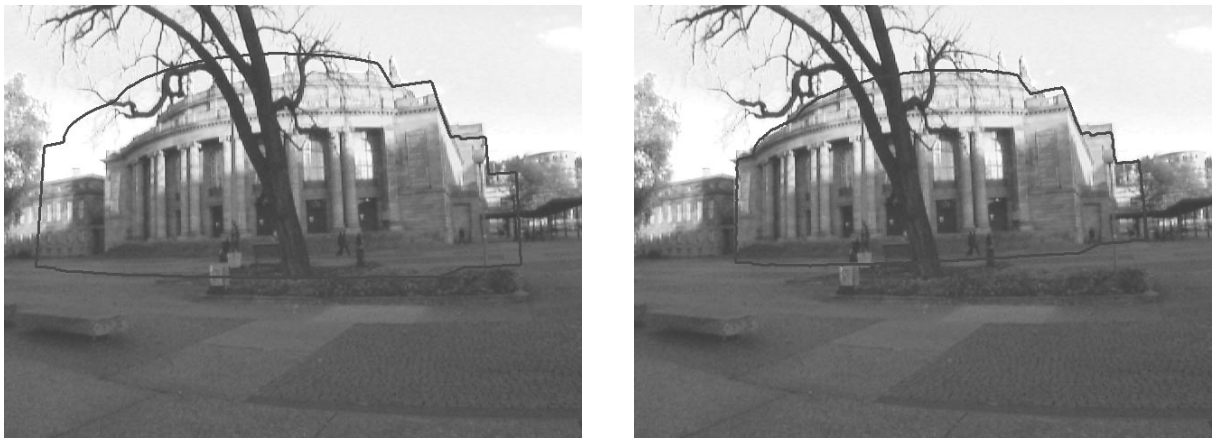

Figure 5: Image of the opera house with building silhouette overlaid. On the left the approximate orientation is used, on the right the result of the detection is displayed. Detection was successful despite the occlusions, shadowing effects and model imperfections.

Figure 6: Image of a museum with building silhouette overlaid. On the left approximate orientation, right after detection.

4.3 Accuracy of Orientation

While the theoretical accuracy of differential GPS is very high, there are many practical limitations. This is especially true when using GPS in built-up areas. Shadowing from high buildings causes poor satellite configurations. At times the signal is lost completely. Additionally signal reflections from buildings nearby causes so called multipath effects further reducing accuracy. As can be seen form Figure 7 on average we experienced an accuracy of 7–10 meters in planar coordinates in our tests. The Z component of the GPS measurement was discarded and substituted by height values from a digital elevation map. However the GPS measurements and the orientation angels obtained from the compass and tilt sensor were sufficiently accurate to serve as approximate values for our automated detection.

Table 1 shows the results corresponding to the images of the opera house in Figure 5. In order to assess the accuracy, the viewpoint was selected at a position well identifiable in a high resolution digital map. The map coordinates thus serve as ground truth for our experiments. The planar coordinates of the users position are displayed as received from the GPS measurement, from the resection after the object recognition and from the map. One can clearly see an improvement of several meters obtained by our approach.

	X	Y	Z
GPS	3513559,82	5404697,37	
Resection	3513551,31	5404698,50	243,97
Map	3513550,25	5404701,23	245,21

Table 1: The position determined by resection has improved over the initial GPS coordinates when compared to ground truth from the map.

In Figure 7 the results of four test cases are summarized. There is a clear improvement in all cases in the accuracy of the planar position. It is interesting to observe that the final accuracy is independent of the initial accuracy of the GPS. The final accuracy is also not the same for all four buildings. This again indicates that the error within the available model data has a strong influence on the result.

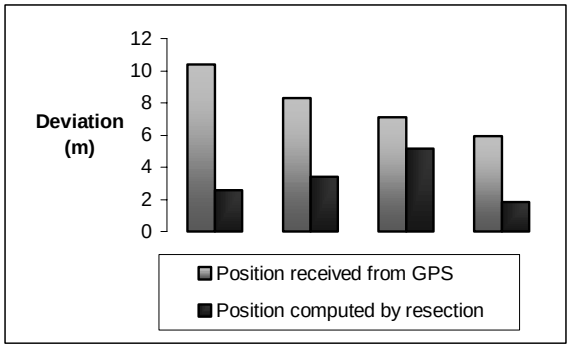

Figure 7: Deviation from ground truth in planar coordinates for four test cases. The deviation has been reduced by our approach in every case.

5. CONCLUSION

We have demonstrated the successful implementation of a fully automated process for the detection of buildings in terrestrial images. The core detection algorithm is based on the Generalized Hough Transform. It has proven to be robust against clutter and occlusion. The silhouette of a building has shown to be a good choice as representation for the shape of the building and that compensated the model imperfections of the 3D city model. The accuracies, which were obtained for the orientation, are sufficient for navigation tasks. Within the article we have shown the use of 3D building models and close-range photogrammetry for location aware applications such as image based orientation and telepointing.

6. REFERENCES

D. H. Ballard, 1981. *Generalizing the Hough transform to detect arbitrary shapes*. Pattern Recognition, 13(2), pp. 111-122.

E. Baltsavias, A. Grün and L. van Gool, 2001. *Automatic Extraction of Man-Made Objects From Aerial and Space Images (III).* A.A. Balkema Publishers.

Fischler, M. A. & Bolles, R. C. 1981. *Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography*. Communications of the ACM, 24(6), pp. 381–393.

Fraser, C. S., 1997. *Digital camera self-calibration*, in: ISPRS Journal of Photogrammetry and Remote Sensing, 52, pp. 149–159.

Fritsch, D., Klinec, D., Volz, S., 2000. NEXUS - Positioning and Data Management Concepts for Location Aware Applications. In: *Proc. of the 2nd International Symposium on Telegeoprocessing* , pp. 171-184.

Grün, A. & Wang, X., 2001. News from CyberCity-Modeler. In: *Automatic Extraction of Man-Made Objects from Aerial and Space Images (III)* , pp. 93-101.

Höllerer, T., Feiner, S., Terauchi, T., Rashid, G., Hallaway, D., 1999. Exploring MARS: developing indoor and outdoor user interfaces to a mobile augmented reality system *Computers & Graphics* **23,** pp. 779-785

Hough, P. V. C. 1962, *Method and means for recognizing complex patterns*. U.S. Patent 3,069,654.

Klinec, D. & Fritsch, D. 2001. *NEXUS – Acquisition of Position Information for Location Aware Applications using Multi Sensors and Mobile Photogrammetry*. Proceedings of ION'01, Salt Lake City, USA.

Koenderink, J.J. & van Doorn, A.J., 1997. The Internal Representation of Solid Shape with Respect to Vision. In: *Biological Cybernetics*, Vol. 32, pp. 211-216.

Kraus, K. 1996. *Photogrammetrie*, 2, Dümmler.

Ulrich, M., Steger, C., Baumgartner, A., Ebner, H., 2001. Real-Time Object Recognition in Digital Images for Industrial

Applications. In *5th Conference on Optical 3D Measurement Techniques*, pp. 308-318.

Wolf, M., 1999. Photogrammetric Data Capture and Calculation for 3D City Models. In: *Photogrammetric Week '99*, pp. 305-313.