METROPOGIS: The next generation

Lukas Zebedin, Andrea Klaus, Barbara Gruber and Konrad Karner

VRVis Research Center.

Inffeldgasse 16/2, Graz, Austria

zebedin@vrvis.at

<http://www.vrvis.at/2d3d>

Remisión Artículo: 16-4-2007

Keywords: Ortho Image Generation, UltraCamD from Vexcel Imaging

Abstract. This paper describes a highly automatic work flow for urban model generation from digital aerial imagery. The images used in our approach feature one high resolution panchromatic channel and four lower resolution channels (red, green, blue and near infrared). In the first step an initial land use classification is performed where a support vector machine is applied to each image. Subsequently the Aerial Triangulation algorithm integrates area- and feature-based points of interest to determine the position and orientation of all images. Afterwards, a digital surface model is generated by a dense image matching procedure. Based on the AT and dense image matching result a true ortho image is computed. Different layers for building blocks, streets, vegetation or water regions are generated from a refined land use classification. The building block layer is used in a subsequent step to extract polygonized buildings.

1. Introduction

Digital aerial cameras are already widely adopted and are continuously substituting conventional aerial cameras. This paper is based upon the UltraCamD from Vexcel Imaging. This particular camera delivers panchromatic images with a resolution of 11500 pixels across-track and 7500 pixels along-track. Additionally it features lower resolution (3680 by 2400 pixels) color and infrared channels. This paper is based upon a data set which was acquired in summer 2005 of the inner city of Graz, Austria. It consists of 155 images flown in 5 strips. The along-track-overlap of this data set is 80%, the across-track overlap is approximately 60%. The ground sampling distance (GSD) is around 8cm. An example of the images used by the described algorithms is depicted in Figure 1.

Figure 1: Input imagery of the UltraCam-D – from left to right the panchromatic, color and infrared channels are depicted.

The paper is structured in the following way: In Section 2 the initial classification is explained, the AT algorithm and subsequent dense image matching are detailed in Section 3 and 4. True ortho image generation and refined classification are dealt with in Section 5 and 6. The urban model generation is described in Section 7. Finally a conclusion and outlook is given in Section 8.

2. Initial Classification

Each input image is classified by a supervised classification algorithm. At the core of the classification there is a support vector machine (SVM). A SVM is a theoretically superior machine learning algorithm which tries to find the optimal boundary between classes which generalize to unseen samples with a minimum error. One benefit of this approach is that the complexity of the machine is only defined by the support vectors, but independent from the dimensionality of the input space. A detailed description of SVM for land use classification is given in [Hsu et al., 2003]. The separability is improved by increasing the dimensionality of the input space which originally consists only of four channels (red, green, blue and infrared) by adding non-linear combinations of those four channels like the normalized difference vegetation index (NDVI) which is the ratio of the green and infrared channel.

Figure 2 illustrates this concept very well – for clarity reasons not all indices are included

Figure 2: This graph illustrates a subset of the features which are used for classification. The NDVI for example is a very salient feature for vegetation even though it just depends on the green and infrared channel. In case of general classes like solid or roof the variation of the signals is much bigger

The SVM has to be trained for each class with a number of representative feature vectors. Often this has only to be done once for the whole flight if the weather conditions do not change too much and the land properties are similar. The result of the initial classification is a primary and secondary probability for each pixel in the input image. This information is later used to create an ortho classification image and to guide the extraction of tie points as unsuitable regions like water and shadow can be excluded.

3. Aerial Triangulation

The basis for the aerial triangulation is a set of correspondences, also called tie points, between neighboring images [Schenk, 1997]. The whole processing chain works completely automatically, therefore a high degree of redundancy in the imagery is desirable because this lends more robustness to the algorithms and introduces very little additional costs as there is no human operator involved.

As a minimum requirement the side overlap should exceed 60% and the forward overlap should surpass 80% - this is important when considering the altitude and velocity of the flight as there is a certain maximum trigger rate of the camera. In the first step of the aerial triangulation the above mentioned tie points are detected using the Harris corner detector and utilizing line intersections [Bauer, 2004].

Finding correct correspondences is approached as an uncalibrated wide baseline problem [Mikolajczyk and Schmid, 2003], therefore a robust descriptor is calculated which is similar to the Sift descriptor used by [Lowe, 1999]. Afterwards the corners are sorted according to their strength. Each feature vector is then matched to find plausible corresponding corners in adjacent images. Each candidate is evaluated using an adaptive area based correlation. Only matches with a highly distinctive score are retained. Further robustness against outliers is added by a back-matching and the constraints imposed by the epipolar geometry which is obtained by applying the well known five point algorithm [Nister, 2003] on samples obtained by the RANSAC algorithm.

As a result the relative orientation for adjacent images is derived. In order to align all cameras in the same coordinate system a scale factor has to be determined for each image pair. This is done using point correspondences for three images.

In the last step of the aerial triangulation a bundle block adjustment optimizes the position and orientation of each camera to minimize the reprojection error and integrates other information like GPS data or ground control points which transform the local coordinate system into a real world coordinate system. Figure 3 shows an oriented block of 155 images.

The sparse reconstruction is used to initialize the dense image matching procedure.

Figure 3: This image shows an oriented block of 155 images (black arrows) of the Graz data set. It contains over 70.000 reconstructed points which are represented by black dots. Each of them was reconstructed using on average 5 corresponding tie points. The mean reprojection error after the block bundle adjustment is below 0.25 pixel.

4. Dense Image Matching

Once the orientation of the cameras is known the dense image matching procedure starts. A good overview of recent advances in this field is given in [Scharstein and Szeliski, 2002]. Because of the steep height discontinuities it has to face, an initialization is required to facilitate convergence of the later steps. This initialization is obtained by a plane sweeping algorithm as proposed by [Collins, 1995]. The approximate range of the possible depth range for the scene is already known from the sparse reconstruction during the AT. In order to get dense depth estimates a plane sweeps that range. All images are projected on that plane and a correlation value is computed for the warped images. Afterwards one depth value is selected for each sampling point from that stack. This depth value, however, does not necessarily need to be the locally best hypothesis, as this search is guided by a semi-local and iterative optimization technique.

This coarse initialization is refined and densified by an iterative and hierarchical algorithm. It starts with the coarse, regular grid produced by the plane sweeping and upgrades its resolution once the optimization is done for the current level.

This optimization consists in modifying the position of the 3D points of the regular grid until the cross correlation of the input images can not be increased any more. If no more improvements are possible additional points are introduced by interpolation between existing grid points, thus a denser regular grid is created. The result of this procedure for the inner city of Graz is shown.

The described algorithm can also deal with large areas as they are split into smaller tiles which are processed independently. This has the additional benefit that it allows for parallel and distributed computation on multiple processors and computers. A detailed and mathematical description of the algorithm can be found in [Zach et. al., 2003].

5. True Ortho Image Generation

The work flow presented in this paper needs ortho photos not only for texturing the digital terrain model (DTM) which is later merged with the polygonal building models to generate a realistic virtual representation of a city, but also to enhance the initial classification.

Our approach to generate the true ortho image is inspired by [Oda et. al., 2004]. The input data for this algorithm is the height field and the orientation of the cameras.

Every point in the height field is then textured from the input imagery. This input imagery can either be panchromatic, color, infrared or even classification images.

The foundation of the algorithm is a hierarchical data structure which permits fast queries of visibility between a point in the height field and a camera position.

Therefore it is possible to calculate in a fast way which cameras are visible from a specific point in the height field. Once the set of cameras is known, the point in the height field is projected into each input image and a candidate value for the texture of the pixel is determined.

Those candidates need to be fused together to obtain the final texture for the respective pixel.

The fusion algorithm, however, depends on the kind of input imagery and should exploit its redundancy to remove obstacles like moving objects. The fusion of panchromatic and infrared values yields the median of all candidate values. Color information is transformed into the CIE LAB color space and clustered there. The mean of the cluster with the best support is used. Classification information finally is fused by a majority voting. The result of this fusion is visible in Figure 4.

Figure 4: Image (a) and (b) are two examples of the input images which are fused together to obtain the true ortho image. There are two moving cars on the street which are removed in the ortho image. Obviously the parking cars can not be removed as they are static objects.

6. Refined Classification

With the help of the ortho images and the height field the initial classification is improved. Especially it helps to split up the solid (streets and roofs) and vegetation (trees, bushes, meadows) classes from the initial classification. As a result separate masks are computed - each of these masks represent one layer indicating where objects of a specific type are. Buildings for example are defined as local height maxima which are restricted to areas classified as solid [Bolter, 2001], roads on the other hand are flat regions of the same class. Vegetation is split up in the same manner using the additional height value to separate trees from flat land.

The building block mask is of special interest because of two reasons: first it is needed to convert the digital surface model (DSM) into a digital terrain model (DTM). This is done by removing all objects from the height field above the ground level. The most important contributions to this type of object is made by buildings and trees. For each connected region of an object of this type a height value is computed by comparing the height inside the region with the average height of the surrounding. Thus by deducing this height from the region the approximate ground level is obtained. Figure 5 demonstrates the transition from DSM to DTM using the classification and color ortho image.

The building block mask is also used during the building block extraction as an initialization for the estimate of the boundary of the buildings.

7. Polygonal Building Generation

The first step of the building reconstruction is the segmentation into separate building blocks. This is straight forward because the building block layer is already available from a previous step. This facilitates also parallel computation because each building block is now treated independently. Using the dense match and the building block mask the building is already

reconstructed as a point cloud. The following algorithms therefore focus on simplifying the representation - the point cloud is a very inefficient representation of planes for example – and improving the facades.

This is important because the facades can not be modelled satisfactorily because it is impossible to represent true vertical walls with the given height field. The first step is to replace the point cloud with geometric shapes to reduce the amount of data needed to represent the building. Currently only planes are used to approximate the building but in future other shapes may also be used like surfaces of revolution.

The replacement algorithm draws random samples from the point cloud which are close together and creates a candidate plane. Afterwards the support for this plane is calculated which is constituted by all points which have a normal distance smaller than a predefined threshold. This support, however, has to be pruned because the points not only have to lie on the same plane but also on the same roof.

Awkward appendices and tentacles are also removed because the support should be a compact cluster of points.

This removal is accomplished by calculating a local density which allows easy identification of isolated points. The clustering is aided by the regular nature of the grid which allows to apply a simple region growing algorithm to this problem. One important restriction during the search for planes is to accept only roof planes with a certain amount of support at the beginning. This threshold is then reduced as more and more planes are found. The reason behind this approach is to find dominant planes first and smaller details later.

This prevents the undesired effect that a dominant plane is partitioned and replaced by various smaller planes which adapt to local noise and possible outliers.

Figure 6: in the upper row the color (left) and classification (right) ortho images are shown. The corresponding height field is shown in the lower left corner. Using the classification a height value is computed for every object above the ground level (trees and buildings) which is deduced from the height field thus generating a DTM.

The second improvement of the point cloud resulting from the dense image matching algorithm concerns the facades. At steep height discontinuities like facades the height field can not represent the true vertical surface. The next step is designed to take care of that problem as the precise position of the facades is estimated.

First the height field is treated as a normal image and used to extract lines. Those lines are the initialization for later optimization phases. We have developed two related ways to align the hypothesis with the real facade.

The first approach is to exploit the fact that a facade is a man made structure and therefore it will be dominated by horizontal and vertical structures. This becomes apparent once the gradients of the texture are calculated and an orientation histogram is computed. If the initialization is close enough to the real facade this histogram will show four distinct peaks. The orientation of the facade is correct if those peaks are at exactly 0, 90, 180 and 270 degrees. By measuring the offset a rotation angle can be computed to make the hypothesis parallel to the real facade. Figure 7 illustrates this concept with a facade from the Graz data set.

Afterwards the position is further modified by optimizing the correlation between images. The drawback of that approach is that the estimated rotation in the first phase is not very exact, but never the less facilitates the convergence of the second phase where the rotation is again optimized.

(a)

Figure 7: **Image (a) shows the texture for the hypothesis which is generated from the extracted line in the height field. The calculated orientation histogram is depicted in image (b). The peaks are obvious but slightly off the expected values (0, 90, 180 and 270 degrees). The required angle to move the peaks towards their expected value can be calculated by intersecting the plain formed by camera center and light green line with a horizontal plain as shown in image (c).**

Because of that drawback we have developed a different method to estimate the rotation of the facade which bases on the assumption that the orientation of the facade can be derived by extracting lines in the aerial images in the vicinity of the hypothesis. Afterwards the facade is aligned on each line and thus the translation is the only remaining degree of freedom because the direction of a line is very precise if there are enough constituting points. Therefore a simple translational sweep of the plane is enough to find a maximum of the cross correlation between two images, thus yielding a optimal position for each extracted line. Finally the hypothesis with the best score is retained.

The fusion of the obtained information is the last step in the generation of the polygonal building block model. The aim of this step is to create a closed polygonal model of each building incorporating all information extracted so far. In order to simplify the task only a 2.5D model is generated so the problem is in fact reduced to 2D. Each detected roof plane is projected into the building block mask which gives an initial segmentation of the building into separate roof planes.

Afterwards the optimized facade lines are also projected into this segmentation. Additionally all close roof planes are intersected to yield even more lines which can aid in the building block generation and guide the following region growing. In order to ensure that the polygonal model is closed, every pixel in the segmentation is assigned to one roof plane. This region growing is guided by the projected lines so that no region can transgress those lines. Finally a segmentation of the building is obtained which partitions the building into regions which can be approximated by a plane.

The next and last step consists in transforming the pixel based segmentation to a vector representation. This is easily done by converting each border between two regions into a sequence of points, thus transforming the segmentation into a graph. In case that there is a projected line nearby the points of these chains are carefully aligned on it, taking care not to cause intersections of chains.

Afterwards each chain is simplified to reduce the number of points. Using a user specifiable threshold points are removed from each chain until the maximum error (normal distance of original points to reduced chain) is exceeded.

Once the segmentation is vectorized and simplified each region is triangulated and converted into a format suitable for display. In order to save space the textures for the polygons are collected in containers of a fixed size (1024 by 1024 pixel for example). Finally the DTM is textured with an ortho image and the buildings are added.

The resulting model can be interactively viewed – Figure 8 gives an impression of the quality of the obtained model.

Figure 8: A screenshot of a virtual flight over Philadelphia.

8. Conclusion and Outlook

This paper describes a work flow which generates polygonal models of urban environments with a minimal human interaction during the supervised classification of the input images. The described algorithms generate a DSM, true ortho images and polygonal approximation of buildings. Future work will focus on enhancing the quality of each step as well as remove the restriction of piecewise planar approximation of buildings which does not work well for domes and spires.

9. Acknowledgements

This work has been done in the VRVis research center, Graz/Austria (<http://www.vrvis.at>), which is partly funded by the Austrian government research program Kplus. We would also like to thank Vexcel Corporation (<http://www.vexcel.com>) for supporting this project.

10. Bibliography

- [Bauer, 2004]** Bauer, J. , Bischof, H., Klaus, A., Karner K., 2004. Robust and fully automated Image Registration using Invariant Features. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 35 (Part B3), 1119-1124.
- [Bolter, 2001]** Bolter, R., 2001. Buildings from SAR: Detection and Reconstruction of Buildings from Multiple View High Resolution Interferometric SAR Data. TU Graz.
- [Collins, 1995]** Collins, R. T., 1995. A Space-Sweep Approach to True Multi-Image Matching. Proc. Conference on Computer Vision and Pattern Recognition, San Francisco, 18-20 June, pp. 358-363.
- [Hsu et al., 2003]** Hsu, C.-W., Chang, C.-C., Lin, C.-J., 2003. A Practical Guide to Support Vector Classification. National Taiwan University, Department of Computer Science and Information Engineering, Taipei 106, Taiwan.
- [Lowe, 1999]** Lowe, D. G., 1999. Object Recognition from Local Scale-Invariant Features. Proc. International Conference on Computer Vision, Corfu, 20-25 September, pp. 1150-1157.
- [Mikolajczyk and Schmid, 2003]** Mikolajczyk, K., Schmid, C., 2003. A performance evaluation of local descriptors. Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Madison, 16-22 June, 2003, pp. 257-263.
- [Nister, 2003]** Nister, D., 2003. An efficient solution to the five-point relative pose problem. Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Madison, 16-22 June, 2003, pp. 195-202.
- [Oda et. al., 2004]** Oda, K., Lu, W., Uchida, O., Doihara, T., 2004. Triangle-based Visibility Analysis and True Ortho Generation. International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences 35 (Part B3), 623-628.
- [Scharstein and Szeliski, 2002]** Scharstein, D., Szeliski, R., 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Computer Vision 47, 7-42.
- [Schenk, 1997]** Schenk, T., 1997. Towards Automatic Aerial Triangulation. ISPRS Journal of Photogrammetry and Remote Sensing 52, 110-121.
- [Zach et. al., 2003]** Zach, C., Klaus, A., Karner, K., 2003. Accurate Dense Stereo Reconstruction using 3D Graphics Hardware. Proc. Eurographics 2003, Granada, 1-6 September, pp. 227-234.