COMPARISON OF PHOTOGRAMMETRIC AND COMPUTER VISION TECHNIQUES - 3D RECONSTRUCTION AND VISUALIZATION OF WARTBURG CASTLE

Helmut Mayer[a], Matthias Mosch[b], Juergen Peipe[c]

[a] Bundeswehr University Munich, D-85577 Neubiberg, Germany - Helmut.Mayer@unibw-muenchen.de
[b] Albert-Ludwigs-Universität Freiburg, D-79106 Freiburg, Germany - matthias.mosch@felis.uni-freiburg.de
[c] Bundeswehr University Munich, D-85577 Neubiberg, Germany - j-k.peipe@unibw-muenchen.de

KEY WORDS: Cultural Heritage, Multimedia, Internet, Calibration, Orientation, Modelling, Comparison

ABSTRACT:

In recent years, the demand for 3D virtual models has significantly increased in areas such as telecommunication, urban planning, and tourism. This paper describes parts of an ongoing project aiming at the creation of a tourist information system for the World Wide Web. The latter comprises multimedia presentations of interesting sites and attractions, for instance, a 3D model of Wartburg Castle in Germany. To generate the 3D model of this castle, different photographic data acquisition devices, i.e., high resolution digital cameras and low resolution camcorder, are used. The paper focuses on the comparison of photogrammetric as well as automatic computer vision methods for camera calibration and image orientation based on the Wartburg data set.

1. THE WARTBURG PROJECT

WARTBURG CASTLE is situated near *Eisenach* in the state of *Thuringia* in Germany. Founded in 1067, the castle complex (Fig. 1) was extended, destroyed, reconstructed, and renovated several times. The Wartburg is a very prominent site with respect to German history, well-known because of the medieval *Contest of Troubadours* around the year 1200 , famous for being the place *Martin Luther* translated the *New Testament* into the German language about 1520, and important as meeting-place and symbol for students and other young people being in opposition to feudalism at 1820. Nowadays, this monument full of historical reminiscences is a significant cultural heritage landmark and, obviously, a touristic centre of attraction.

Figure 1. Wartburg Castle

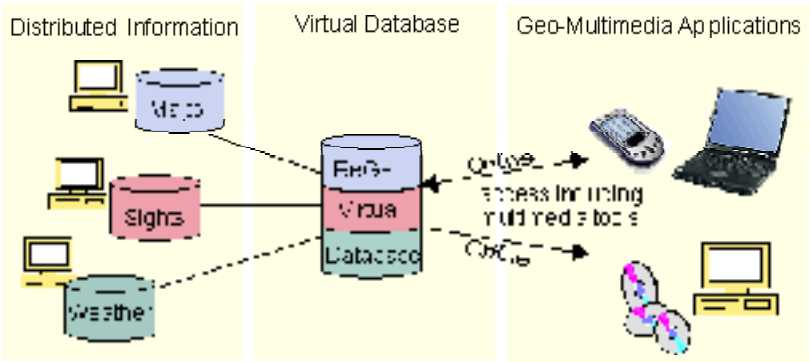

Figure 2. ReGeo system (Frech & Koch, 2003)

There is an ongoing research project funded by the *European Union* aiming at the development of a comprehensive online tourist information system of the *Thuringian Forest* area, based on a geo-multimedia database. The project ReGeo ("Multimedia Geo-Information for E-Communities in Rural Areas with Eco-Tourism" - Frech & Koch, 2003) offers tourists essential and useful information about their holiday region, but also supports local business and administrators (Fig. 2).

The touristic infrastructure and sights can be explored by map-guided quests as well as by alphanumeric thematic search. For geographical visualization, 2D tourist maps, aerial images together with vector data as well as 3D models and 3D sceneries are provided. To generate a 3D model of Wartburg Castle being a highlight of touristic interest in the *Thuringian Forest*, photogrammetric data acquisition and 3D object reconstruction have been carried out.

Examples of 3D modeling and visualizing architectural objects can be found in the literature related to photogrammetry, surveying as well as computer vision. The "photogrammetric way" including a detailed and precise 3D object reconstruction is explained, e.g., in Hanke & Oberschneider (2002), Daskalopoulos et al. (2003), whilst computer vision techniques are emphasized, e.g., in Pollefeys et al. (2003), El-Hakim (2002).

The different approaches have their advantages and limitations. This paper is intended to compare photogrammetric and computer vision methods for camera calibration and image orientation based on parts of the Wartburg data set.

2. OBJECT RECORDING

The Wartburg is built on the top of a rocky hill. The terrain slopes steeply away on all four sides of the castle complex. Access to the inner courtyards is possible solely via a small drawbridge. The outer facades of the Wartburg could be photographed from only few viewpoints. Images taken from an ultra-light airplane had to be added to get a sufficient ray intersection geometry for the photogrammetric point determination. The inner courtyards were successfully recorded from the two towers, some windows, and from the ground. Three image acquisition devices were used: The Rollei d30 metric[5] and Canon EOS D60 SLR cameras, and a Sony DCR-TRV738E camcorder (cf. Tab. 1 for tech. specifications).

The 5 megapixel Rollei camera provides some features of a metric camera, such as fixed focal length and a rigid connection between lens and CCD chip inside the camera. In addition, two focusing stops can be fixed electronically, i.e., the interior orientation parameters at these two focus settings can be regarded as known over a period of time, if they were once determined by camera calibration. The d30 metric[5] used within the Wartburg project was calibrated for the focal length f_1 = 15 mm and f_2 = 30 mm. The setting f_2 was used for the images taken from the ultra-light airplane, whilst f_1 was employed for all other images. The Canon EOS D60 providing a wide angle lens, 6 megapixel sensor and larger image size as the Rollei camera was used to record several parts of the courtyards which could be acquired only from a shorter distance. The camcorder served in addition to collect some overviews and short image sequences (movies) to be presented in the tourist information system.

	Rollei d30 metric[5]	Canon EOS D60	Sony DCR-TRV738E
Number of pixels	2552 x 1920	3072 x 2048	720 x 576
Sensor format	9 mm x 7 mm	22 mm x 15 mm	
Lens (focal length)	10 mm - 30 mm (= 40 mm - 120 mm for 35 mm camera)	20 mm (SLR) (= 32 mm for 35 mm camera)	3.6 mm - 54 mm (= 48 mm - 720 mm für 35 mm camera)
Image data	6.4 MB uncompressed raw data per image	7.4 MB uncompressed raw data per image	0.8 MB

Table 1. Technical Specifications of the cameras used in the project

3. PHOTOGRAMMETRIC RESTITUTION

A subset of 10 Rollei images covering a part of the first courtyard was selected to compare the 3-D reconstruction process normally used in close-range photogrammetry with methods preferably applied in computer vision. The digital images were taken parallel to the object facade having relatively short distances between the camera positions. Image data were obtained by manual pointwise measurement within the *AICON DPA-Pro* and *PhotoModeler* software. Then, interior and exterior orientation parameters as well as 3-D coordinates of object points were determined by self-calibrating bundle adjustment. Fig. 3 shows the result of the visualization of this part of the castle.

Figure 3. Wartburg Castle: Part of the first courtyard

4. FULLY AUTOMATIC SEQUENCE ORIENTATION, AUTO-CALIBRATION, AND 3D RECONSTRUCTION

The goal of this part was to investigate, whether a sequence of images, for which the only thing known is, that they are perspective and that they mutually overlap, is enough for a metric reconstruction of the scene. Additionally we were interested to compare the automatically computed camera calibration information to given calibration information.

4.1 Sequence Orientation Based on the Trifocal Tensor

While other approaches use image pairs as their basic building block (Pollefeys, 2002), our solution for the fully automatic orientation of an image sequence relies on triplets which are linked together (Hao and Mayer, 2003). To deal with the complexity of larger images, image pyramids are employed. By using the whole image as search-space, the approach works without parameter adjustment for a large number of different types of scenes.

The basic problem for the fully automatic computation of the orientation of images of an image sequence is the determination of (correct) correspondences. We tackle this problem by using point features and by sorting out valid correspondences employing the redundancy in image triplets. Particularly, we make use of the trifocal tensor (Hartley and Zisserman, 2000) and RANSAC (random sample consensus; Fischler and Bolles, 1981). Like the fundamental matrix for image pairs, the trifocal tensor comprises a linear means for the description of the relation of three perspective images. Only by the linearity it becomes feasible to obtain a solution when no approximate values are given. RANSAC, on the other hand, gives means to find a solution when many blunders exist.

Practically, first points are extracted with the Förstner operator. In the first image the number of points is reduced by regional non-maximum suppression. The points are then matched by (normalized) cross-correlation and sub-pixel precise coordinates are obtained by least squares matching. To cope with the computational complexity of larger images, we employ image pyramids. On the coarsest level of the image pyramid, with a size of approximately 100 x 100 pixels, we use the whole image size as search space and determine fundamental matrices for image pairs. From the fundamental matrices, epipolar lines are computed. They reduce the search space on the next level. There, the trifocal tensor is determined. With it a point given in two images can be projected into a third image, allowing to check a triple of matches, i.e., to sort out blunders. For large images, the trifocal tensor is also computed for the third coarsest level. To achieve highly precise and reliable results, after the linear solution projection matrices are determined and with them a robust bundle adjustment is computed for the pairs as well as for the triplets.

To orient the whole sequence, the triplets are linked. This is done in two steps. First, the image points in the second and third image of the nth triplet are projected into the third image of the n plus first triplet by the known trifocal tensor for the n plus first triplet. As the (projective) 3D coordinates of the nth triplet are known, the orientation of the third image in the projective space of the nth triplet can be computed via inverse projection. To obtain high precision, a robust bundle adjustment is employed. In the second step, 3D coordinates in the coordinate system defined by the nth triplet are determined linearly for all points in the n plus first triplet that have not been computed before. The solution is again improved by robust bundle adjustment. Starting with the first image, this incrementally results into the projective projection matrices for all images as well as in 3D points. After having basically oriented the sequence on the two or three coarsest levels of the image pyramid, finally, the 3D points are projected into all images via the computed projection matrices. The resulting points are then tracked over one or two levels through the pyramid.

Figure 4 gives results for the first four images of the sequence taken with the Rollei d30 metric[5] camera. One can see that the points have been tracked pretty well even for the wall close to the camera, where the disparities are rather large. For the whole sequence of ten images we have obtained 56 10-fold, 41 9-fold, 91 8-fold, 71 7-fold, 55 6-fold, 30 5-fold, 41 4-fold, and 44 3-fold matches after robust adjustment with a standard deviation of 0.08 pixels.

Figure 4. The first four images of a sequence taken with the Rollei d30 metric[5] camera and the points tracked through it.

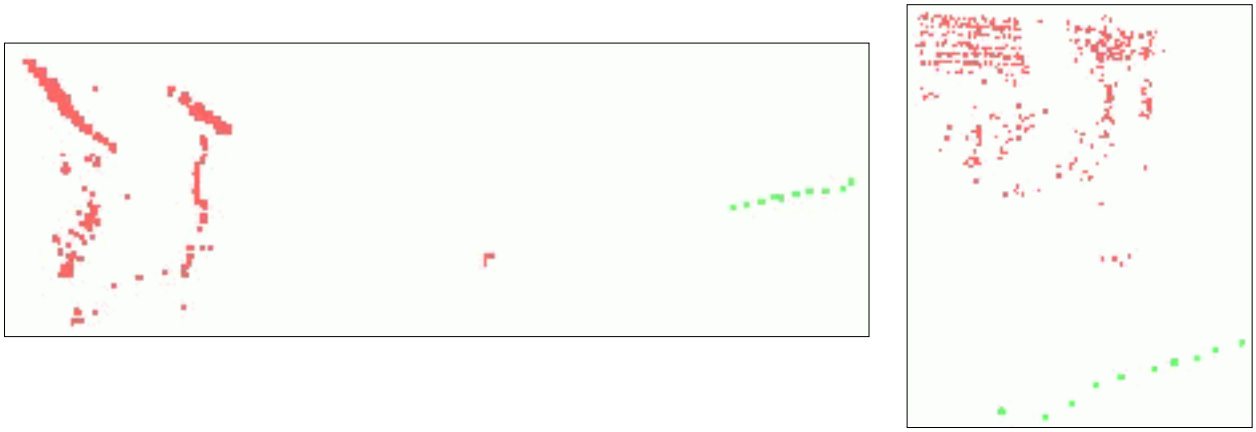

Figure 5. Two views of the VRML model generated fully automatically from the sequence. The cameras are given in green.

4.2 (Auto-)Calibration of the Sequence

Basically, to obtain a metric, i.e., calibrated sequence, the orientations as well as the 3D points have to be upgraded to metric space. Here, we first assume, that the calibration matrix is already given. In a second step we explain, how we obtain an approximation for the calibration.

To calibrate a sequence given a calibration matrix containing the ratios of the principal distance to the image width and height, the principal coordinates in terms of the image width and height, as well as the sheer of the sensor (Hartley and Zisserman, 2000), we use a linear algorithm for the image pair. It results in orientations for the two images as well as in (metric) 3D points. The 3D points are used to calculate the orientations for succeeding images which in turn can be used to compute 3D coordinates for points not visible in the first two images. (Metric) bundle adjustment is employed to determine a highly precise and reliable result.

To obtain a solution also if no calibration information is available, we have developed a solution based on sampling, which gives reliable results in most cases, as long as the principal point is not too far from the center of the image. Though we are therefore not able to deal with parts cut out of larger images, the solution is from our experience (Hao and Mayer, 2003) more stable, especially for triplets or short sequences, than those based for instance on the absolute conic (Hartley and Zisserman, 2000). The basic idea is to set the sheer as well as the principal point to zero. Then, the size of a pixel is assumed to be a square and ratios of principal distance to image width from 2.5 to 0.5 are sampled in 30 logarithmic steps. For all steps the first triplet is calibrated using the resulting calibration matrices. The criterion for choosing the ratio of principal distance to image width is the average standard deviation of the image coordinates. After obtaining the principal distance for the width, the ratio of the width and the height of the pixel is varied from 1.2 to 0.8, i.e., the principal distance for the image height is computed. Finally, more precise values for the principal distances are determined by a bundle adjustment with additional parameters.

For the above sequence we first used the given calibration information. The ratio of principal distance to the image width is known to be 1.7219, the ratio to the image height is 2.2887, and the principal point is close to the center of the image (-0.0112, -0.0298). The skew is assumed to be zero. With this information we have obtained a standard deviation of the calibrated sequence of 0.39 pixels. We have produced a VRML (the virtual reality modeling language; ISO / IEC 14772-1:1997 standard) model from the calibrated 3D points. Two views with the cameras in green and the points in red are given in Figure 5. Our procedure for auto-calibration resulted in 1.84 for the ratio of the principal distance to the image width and 2.34 for the ratio to the image height. The standard deviation in image space was estimated to be 0.48 pixels. This is not an extremely good result, but for visualization the differences were minor. One also has to consider that this is a relatively challenging scene, as the basis between the single cameras is rather small compared to the distance to the object.

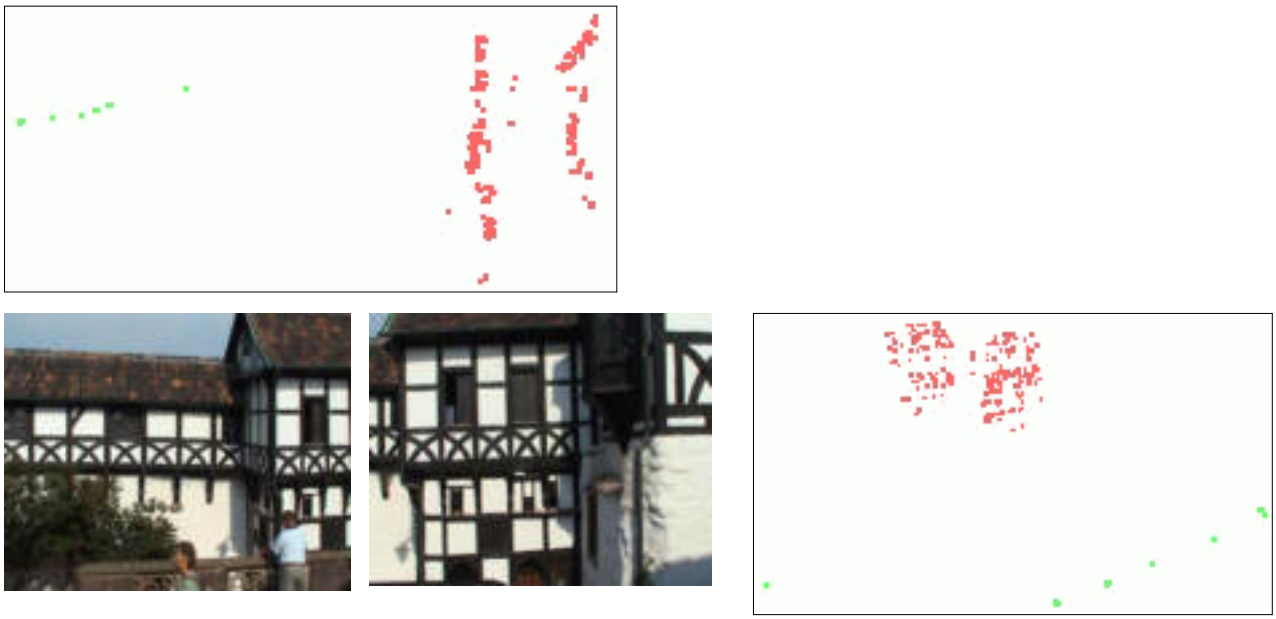

Figure 6. First and last image of an image sequence taken with the Sony camcorder (left below) and two views of the VRML model fully automatically generated from the sequence (top and right).

Figure 6 shows the first and the last image of a sequence taken with the Sony camcorder and two views of the resulting VRML model. Of the 1500 images we have taken only seven to minimize the computational effort. Overall, we obtained 71 7-fold, 60 6-fold, 55 5-fold 24 4-fold, and 4 3-fold points after robust adjustment and an error of 0.06 pixels. For this camera only the range of the focal length is known, but not the pixel size or the metric size of the sensor, i.e., the parameters of the camera matrix are unknown. Our auto-calibration scheme resulted into 1.66 for the ratio of principal distance to image width and 2.38 for the ratio of principal distance to image height. The precision in image space was 0.25 pixels. It should be noted that the results have been obtained without user interaction and with the same set of parameters as for the above sequence. The two views of the VRML model generated for the points and the cameras shows that the metric geometry of the scene has been recovered fairly well.

4.3 Structure computation

Our approach for the computation of 3D structure (Mayer, 2003), i.e., a disparity image in the first place, from an image pair or triplet is based on the algorithm for cooperative disparity estimation proposed in Zitnick and Kanade (2000). The respective image pair is resampled along the epipolar lines. Matching scores are computed by cross correlation and absolute differences and written in a 3D array made up of image width, height, and disparity range. As the computational effort depends directly on the range of the disparity, we compute this range by projecting the points reliably determined for the image triplet onto the epipolar lines.

Figure 7: Disparity map (left) and visualizations based on the disparity map

The basic idea of Zitnick and Kanade (2000) is a cooperation between support and inhibition. Inhibition enforces the uniqueness of a match. Assuming opaque and diffuse-reflecting surfaces, a ray of view emanating from a camera will hit the scene at one point only. The idea is to gradually weight down all matches on a ray besides the strongest. Support is realized by filtering the 3D array by a 3D box filter. By this means all matching scores corroborate locally to generate a continuous surface. We have improved this scheme with different means, most notably a combination of image and disparity gradient to avoid smoothing away details and a detailed treatment of occlusions. Additionally, we use the information of a third image by projecting the results of the cooperative computation into the third image by means of the trifocal tensor and by modification of the 3D array based on the correlation score of the first and the third image.

Figure 7 gives the result for the computation of the disparity map from three images of the image sequence given above. The disparity map outlines the important structures of the scene. While moving from left to right as done in the first two visualizations results into the movement of the wall in front of the camera and an attractive view, a movement upwards as in the rightmost image shows problems from non-modeled occlusions.

4.4 Analysis of the Automatic Approach

We have shown a way to fully automatically generate metric 3D structure from the weak information of the perspective images of a sequence. Yet, there is ample room for improvement before this approach will become practically relevant. Most notably, it is necessary to link the results for structure computation in two or three images, to obtain a coherent result for the sequence. Then, there is a host of robustness issues which has to be solved before all types of sequences can be handled reliably. Finally, in the foreseeable future, the automatic approach will have problems, when the perspective skew between images is too large, i.e., when the angle between consecutive images is too large. This can be avoided by taking more images. This is disadvantageous when taking the images, but it helps to avoid a lot of manual work.

5. CONCLUDING REMARKS

Image-based survey and 3D modeling of architectural objects can be performed by photogrammetric as well as computer vision and computer graphics methods. The latter allow fully automatic feature extraction, orientation, and 3D object reconstruction, even if the knowledge about the geometry of the image sequence is weak. Auto-calibration is possible in case that the interior orientation data have changed during the image recording (focusing, zoom) or are not available at all. Preferably, photogrammetric and computer vision methods are integrated to generate virtual models from images as precise and reliable as required.

It should be noted that the results presented in this paper are preliminary. The reconstruction and modeling of Wartburg Castle is still in progress.

6. REFERENCES

Daskalopoulos, A., Georgopoulos, A., Ioannidis, Ch., Makris, G.N., 2003. A Trip from Object to Image and Back. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Ancona / Italy, Vol. XXXIV (5/W12), pp. 141-144.

El-Hakim, S.F., 2002. Semi-Automatic 3D Reconstruction of Occluded and Unmarked Surfaces from Widely Separated Views. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Corfu / Greece, Vol. XXXIV (5), pp. 143-148.

Fischler, M., Bolles, R., 1981. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM,* 24 (6), pp. 381-395.

Frech, I., Koch, B., 2003. Multimedia Geoinformation in Rural Areas with Eco-Tourism: The ReGeo-System. In: *Information and Communication Technologies in Tourism 2003*, Springer, Vienna / New York, pp. 421-429.

Hanke, K., Oberschneider, M., 2002. The Medieval Fortress Kufstein, Austria - An Example for the Restitution and Visualization of Cultural Heritage. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Corfu / Greece, Vol. XXXIV (5), pp. 530-533.

Hao, X., Mayer, H., 2003. Orientation and Auto-Calibration of Image Triplets and Sequences. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Munich / Germany, Vol. XXXIV (3/W8).

Hartley, R., Zisserman, A., 2000. *Multiple View Geometry in Computer Vision*. Cambridge University Press, Cambridge, UK.

Mayer, H., 2003. Analysis of Means to Improve Cooperative Disparity Estimation. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Munich / Germany, Vol. XXXIV (3/W8).

Pollefeys, M., Verbiest, F., van Gool, L., 2002. Surviving Dominant Planes in Uncalibrated Structure and Motion Recovery. In*: Seventh European Conference on Computer Vision*, Vol. II, pp. 837-851.

Pollefeys, M., van Gool, L., Vergauwen, M., Cornelis, K., Verbiest, F., Tops, J., 2003. 3D Capture of Archaeology and Architecture with a Hand-Held Camera. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Ancona / Italy, Vol. XXXIV (5/W12), pp. 262-267.

Zitnick, C., Kanade, T., 2000. A Cooperative Algorithm for Stereo Matching and Occlusion Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22 (7), pp. 675-684.