Multi-Sensors and Multiray Reconstruction for Digital Preservation

DIETER FRITSCH, ALI MOHAMMAD KHOSRAVANI, ALESSANDRO CEFALU, KONRAD WENZEL, Stuttgart

ABSTRACT

The use of multi-sensor systems is demonstrated in this paper. Firstly, the potential of modern smart phones is used to reconstruct 3D shapes for Cultural Heritage applications. Secondly, we let collect the Microsoft Kinect sensor system dense point clouds for 3D indoor modelling. Thirdly, a multi-camera system is mounted onto a rectangular aluminium frame to collect thousands of images of two reliefs, which are processed by refined image matching algorithms. The resulting point clouds are of very dense resolution and high accuracy. The potential of the implemented image matching algorithm seems to be very flexible, fast and lean with regard to storage requirements. Three examples demonstrate the application of the image processing pipeline.

1. INTRODUCTION

The use of multi-sensor systems is trendy in the fields of computer science, computer vision, machinery inspections and many others. Also photogrammetry is using multi-sensor systems since quite a long time. Firstly, additional flying height observations have been used for external height controls (in the 1930s), then GPS positions of airborne camera locations have been measured and introduced as weak datum observations in bundle block adjustments (1980s). In the 1990s integrated GPS/INS observations have approved the potential of 3D point determination without ground control. Image matching has been invented to overcome the manual image coordinate measurements (1980s). This method has been used for automated aerial triangulation and automated DTM generation for the last 20 years. Summarizing the developments in airborne photogrammetry, it is fair to say that the knowledge of advanced bundle block adjustments and the application of image matching algorithms belongs to the toolkit of digital photogrammetry since many years.

With the widespread use of smart phones their capabilities as an integrated sensor system has been proven from the beginning. Today, many smartphones have a CCD sensor, GPS receiver, magnetic sensor and acceleration sensors on board, besides the functionality of communication and online data traffic exchange.

To use a smart phone as a photogrammetric data collection system is still a challenge. Firstly, the individual parameters of its several sensors have to be read-out and calibrated. Secondly, some knowledge for using those parameters in a straightforward approach is necessary. But at the end it seems to be a fact that the proper application of all those parameters in a rigorous photogrammetric approach delivers 3D object points to be used for some documentation purposes. The advantages of using smart phones for 3D documentation purposes are obvious: It is always available, of low-cost and delivers many sensor data.

With the progress in the field of image matching, e.g. Semi Global Matching (H. Hirschmüller, 2008) close range imaging sensors can be used to compute very dense point clouds. Also those sensors are low-cost and can be combined to multiple units providing not only the necessary stereoscopic overlaps but also multiray geometry, and last but not least, redundancy to improve the accuracy considerably.

In a case study the potential of smart phones, the use of the Microsoft Kinect controller and a close-range photogrammetric system consisting of a multi-camera system and advanced matching algorithms is demonstrated. The first two are delivering 3D object points and point clouds to be used for simple tasks in the field of Cultural Heritage. Professional digital preservation needs more

sophisticated sensors or a combination of low-cost sensors with a professional pipeline of image processing, which is addressed in the following section.

1.1. High Definition Digital Preservation

Digital documentation of 3D objects is an important task for the preservation of historical monuments and to get up-to-date information. Building Information Models (BIM) provide a basic data set to be used for many purposes, ranging from "as-built" documentation over rendering "as-built with planned" to virtual walk-troughs using all real and virtual data sets of the building. Quite often, High Definition Surveying (HDS) is used to provide dense point clouds for further refinement and processing. Although HDS seems to be efficient and powerful and is an emerging technology, sometimes its application is limited especially when viewing the 3D object under unfavourable angles and distances. Moreover, its resolution might be not sufficient if archaeological statues, reliefs and 3D jigsaws have to be sampled very densely.
One method to overcome this problem is to complement HDS point clouds with selected dense point clouds collected with an industrial 3D scanner (e.g. Minolta Vivid 910, Leica Geosystems T Scan). This is reported to be efficient but very time-consuming and sometimes costly (G. Sansoni et al., 2009). The reason to use 3D scanners in the cultural heritage field is simple: They meet the requirements as there is always a challenge in precision, dense sampling and the integrity of the object.
Surface reconstruction from imagery represents an efficient method of accurate 3D data acquisition since imagery from digital cameras provides a high density of information, represented by geometric and radiometric resolution. Therefore, solutions for 3D surface reconstruction from imagery are particularly suitable for close range cultural heritage applications, where the requirements regarding acquisition efficiency, flexibility, but also spatial resolution and precision are high.
In March 2011 the ifp Research Group: "Close Range Photogrammetry and Terrestrial Positioning Systems" has got an industrial contract to collect photos for a very dense 3D point cloud generation of the two Tympana of the Royal Palace in Amsterdam. The work took place as part of the restoration work which is currently carried out at the whole building. For this purpose we planned to use a multi-camera system incorporating a dense matching implementation (see section 5). Regarding the hardware, one goal for the Amsterdam project was to combine low to medium cost hardware components to set-up a more sophisticated multi-sensor system. At the same time the system design had to match many other conditions. One main challenge was to meet the dense sampling and accuracy criteria
The tympana are located at a height of about 30m and were reachable via a scaffold which surrounded the building. The Tympana have a triangular shape of about 25m in width and about 6m in height. In order to get access in height three scaffold levels were provided. The scaffold itself offered a maximum working distance of about 90cm, often less. Its' floor was of wooden planks and additionally the scaffold construction tended to swing, for example when hit by wind or when the construction site elevators were used.

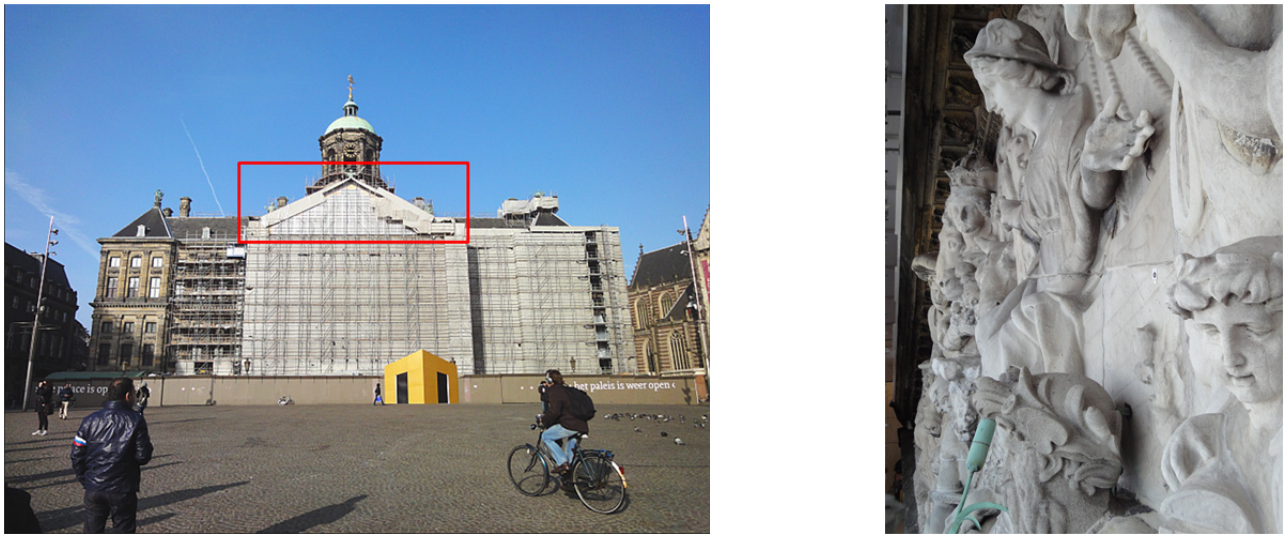

Figure 1: The Amsterdam Tympana (Left: Location. Right: Photo with details)

The final goal was to provide a very dense point cloud using refined image matching for both tympana, with a sampling and accuracy of less than 1mm. Parallel to the image data collection both façades have been scanned by HDS for providing the geospatial reference frame to merge the tympana with the HDS point cloud later on.

2. SMARTPHONES AS MULTI-SENSOR SYSTEMS

There is a growing effort in improvement of smartphones regarding computation capabilities, connectivity and sensor integration. In this section, some capabilities and limitations of the integrated sensors in one exemplary smartphone regarding the usage of provided images, position and orientation data, in a simple photogrammetric application has been investigated.

2.1. The Smartphone Specifications used in this Study

The smartphone used in this project, HTC Hero, employs the Android open source operating system. Using the Java platform enables the full control of the integrated sensors. This smartphone employs a 5 Megapixels digital camera, a GPS receiver, a 3D accelerometer and a 3D digital compass. The camera was calibrated using a suitable 2D calibration field based on the well-known Brown model, in a free net bundle block adjustment (D.C. Brown, 1971). The device's position and orientation can be derived from the output of the GPS receiver, 3D accelerometer and 3D digital compass, respectively.

2.2. Accuracy of the Integrated Sensors

GPS: The accuracy of single point positioning by GPS is about 15-25m (95%). Having redundant satellite data, the accuracy of single point positioning can be further improved. Moreover, if a set of coordinates are measured in a relatively short time interval, due to relatively slow changes in the satellites constellation, and also since all the points receive the satellite signals from a same ionospheric patch, the relative positioning accuracy would be improved dramatically (P.K. Enge et al., 1988; A. El-Rabbany, 2002). The new generation of hand-held GPS receivers is enabled to use WAAS and EGNOS services which provide users with ionospheric corrections (FAA and ESA websites, 2010). It reduces therefore the main GPS error source.

Accelerometer and Digital Compass: When the device is not moving, the only acceleration is caused by the gravity. Therefore one can compute the orientation angles using the magnitudes of the acceleration components on the device's 3 axes and the digital compass data. To have a feeling about the standard deviations of the orientation angles, the device was placed in a horizontal, then a vertical and finally an arbitrary state while collecting these data (with at least 20 samples per case). Results are given in table 1.

Angle	Mode	Average	Std. Dev.
Azimuth	Horizontal	-160.0°	1.0°
Azimuth	Vertical	254.0°	8.0°
Azimuth	Arbitrary	89.6°	1.3°
Pitch	Horizontal	4.5°	0.2°
Pitch	Vertical	-87.7°	0.3°
Pitch	Arbitrary	-40.0°	0.3°
Roll	Horizontal	-0.8°	0.1°
Roll	Vertical	57.7°	7.7°
Roll	Arbitrary	0.0°	0.2°

Table 1: Empirical standard deviation of the orientation angles

2.3. Sensors' Data Focused as the Exterior Orientation Parameters

To have an impression about the usage of the integrated sensors, the sensors' data were contributed in a bundle block adjustment, as approximate values of the exterior orientation parameters (also called 'weak' datum), in a relatively small example.

Figure 2: The gate tower of the Hirsau abbey

The measured object is the gate tower of the Hirsau abbey, close to the city of Calw, in south-west of Germany (figure 1). The object has also been measured by HDS. Thus the accuracy of the 3D reconstruction by the presented approach can be evaluated by a comparison between the CAD model reconstructed using this approach and the one resulting from the HDS point cloud.

The initial values of the object coordinates can be computed either by a forward intersection having the sensors' data as the approximation of the exterior orientation parameters, or by a relative/absolute orientation. However, in this example, only the approximate values computed by the latter method converged the adjustment, because of existing relatively high amount of outliers in the observations of the azimuth angle. The bundle block adjustment was implemented by two approaches.

Firstly the adjustment was implemented using a free net adjustment (D. Fritsch and B. Schaffrin, 1981). The image coordinates were the observations only. The observations of the camera position and orientation were used as the approximation of unknown exterior orientation parameters. This prevents the shape of the network to be influenced by possible inconsistencies between the datum observations (the camera positions and orientations). In this case, the network is fitted to the approximate values of the unknown parameters.

Secondly, an over-determined bundle block adjustment was set up by adding the exterior orientation observations to the system of equations. Regarding the Helmert model for the variance components estimation (E.W. Grafarend et al., 1980), standard deviations were estimated for each set of the

observations (table 2). The results show that one can expect an accuracy of about 5°-6.5° for the camera orientations and 0.3-0.5m for the relative camera positions (after removing the outliers from the adjustment process).
The reason for estimating such a small standard deviation factor for the image coordinates is the magnitude of the image coordinates residuals, which are unusually small. Small residuals are not always desirable, since it might be a sign of existence of poorly controlled parts in the network. The analysis of the local redundancy numbers verifies this assumption (A. Leick, 2004); some of the local redundancy numbers are near to zero. This rather weak geometry is because of the thin shape of the object in combination with the ring configuration of the images block. It can be improved by efficiently increasing the number of images.

Parameter	Std. Dev.	Estimated Std. Dev. Factor
Image coordinates	1 pixel	0.4
Azimuth	6.5°	1.0
Pitch and Roll	5.2°	1.0
Easting and Northing	0.50m	1.0
Height	0.36m	1.0

Table 2: Estimated standard deviations for each set of the observations

2.4. Verification of the 3D Model

The absolute accuracy of the 3D reconstructed model can be investigated by comparing the photogrammetrically determined points or distances with respective reference values (control data) measured independently with a higher accuracy. For this reason, some lengths on the 3D model created by the presented approach are compared with the corresponding lengths on a 3D model resulted by HDS data (table 3). The measured lengths are the widths of the façades of the reconstructed object, which is an octagonal cylinder (figure 2).
Results show that the model resulted from the free net adjustment is around 2% (85% significance) larger than the model created by HDS data. For the over-determined system, the model is around 3% smaller (96% significance) than the reference. Of course, the scale of the model is directly related to the quality of GPS observations, and the ability to detect possible outliers in these observations.

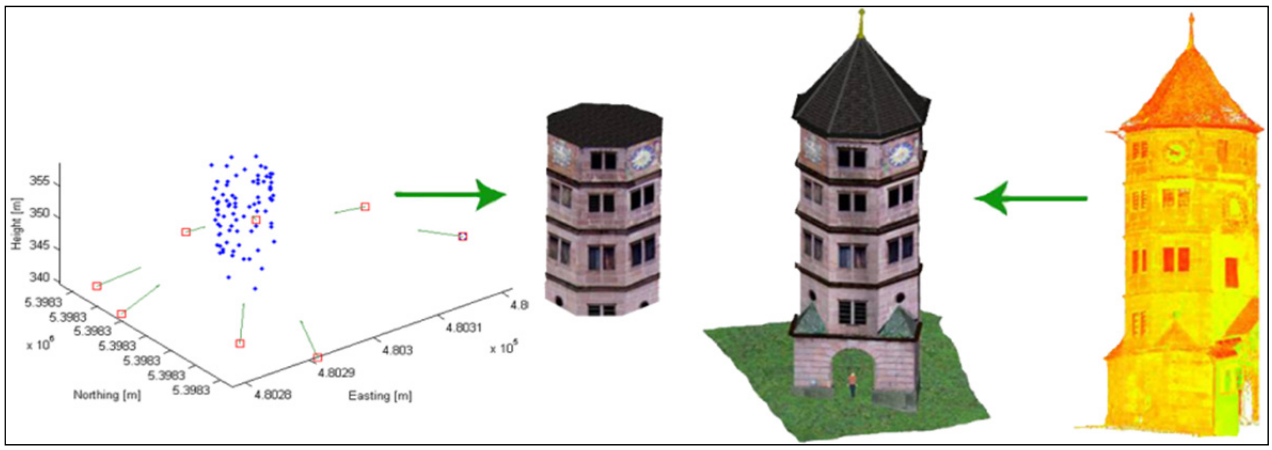

Figure 3: 3D models created by the described photogrammetric method (left) and terrestrial laser scanner (right)

	Free net adjustment				Over-determined system			
	L_1[m]	L_2[m]	Scale= L_1/L_2	$L_2.Scale_{Avg}$-L_1[m]	L_1[m]	L_2[m]	Scale= L_1/L_2	$L_2.Scale_{Avg}$-L_1[m]
	2.956	2.886	1.024	-0.011	2.767	2.886	0.959	0.027
	2.936	2.850	1.030	-0.028	2.813	2.850	0.987	-0.054
	3.036	2.984	1.017	0.009	2.845	2.984	0.953	0.044
	2.973	2.909	1.022	-0.005	2.820	2.909	0.969	-0.004
	3.025	2.922	1.035	-0.044	2.875	2.922	0.984	-0.046
	3.045	2.956	1.030	-0.029	2.873	2.956	0.972	-0.011
	2.916	2.910	1.002	0.053	2.793	2.910	0.960	0.024
	2.984	2.980	1.001	0.057	2.864	2.980	0.961	0.021
Avg.	N/A	N/A	1.020	0.006	N/A	N/A	0.968	0.000
Std. Dev.	N/A	N/A	0.013	0.037	N/A	N/A	0.012	0.036

Table 3: Verification of the 3D model, L_1 is measured length and L_2 is reference length

This example shows, that the accuracy bounds of the integrated sensors are generally safe enough for the convergence of a classical bundle block adjustment, except the digital compass, which can be easily disturbed by environmental magnetic fields. The individual sensors integrated in smartphones also might be used for supporting computer vision and other applications like indoor navigation.

3. POINT CLOUD COLLECTION BY MICROSOFT KINECT SYSTEM

In recent years, active sensing is widely needed in many indoor applications like 3D indoor modeling, mobile mapping, etc. This can be provided by the usage of laser scanners (expensive and relatively too large), or time of flight cameras. A low-cost alternative can be the Microsoft Kinect system which is originally developed as a user interface for the Xbox 360 game console.

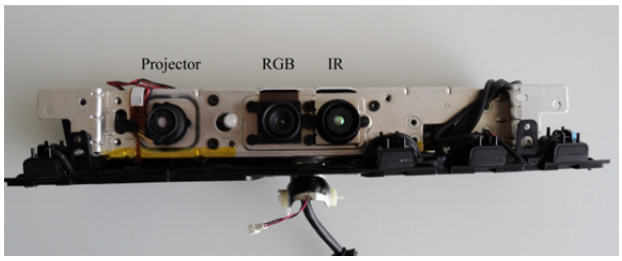

Figure 4: Disassembled Kinect

As depicted in figure 4, this system consists of an infrared (IR) laser projector, a monochrome IR and an RGB CMOS sensor. The system projects a structured IR speckle pattern on the object surface. The pattern is then collected by the IR camera in 30 Hz at VGA resolution (640×480 pixels) with 11-bit radiometric depth. In order to compute range images, automatic stereo measurement is realized by the analysis of the collected pattern.

When used with the Xbox software, the Kinect has a practical ranging limit of 1.2-3.5m, although the sensor maintains tracking with an extended range of approximately 0.7-6 m. The system has 57° horizontally and 43° vertically angular field of view (Wikipedia-Kinect, 2011). Due to internal errors and the errors in the IR pattern matching, at an exemplary distance of 2.6m one can expect an RMS of about 12mm and a maximum error of about 90mm for the noise of the resulted point cloud (N. Haala et al., 2011).

3.1. Point Cloud Texturing and Alignment

The color and depth images of the system can be accessed using software libraries and software development kits (SDKs). In order to texture the point cloud, the Kinect stereo camera system was

calibrated at a suitable calibration field. Image blocks of RGB and IR cameras were used in a photogrammetric bundle block adjustment using the Australis software. Therefore, for each pixel in the range image a corresponding value from the RGB image can be interpolated, while simultaneously the effect of the lens distortion can be removed. Figure 5(a) depicts a textured point cloud resulted from a single frame capture of an exemplary room.
In data acquisition, in order to cover a whole room space, multiple data collection and alignment is needed. In this example, the multiple point clouds were aligned by automatic feature extraction and matching between the consecutive RGB photos, and making use of the estimated relative orientation between the RGB and IR cameras. In other words, in consecutive RGB images corresponding points shall be determined by e.g. a SIFT feature extraction and matching (D.J. Lowe, 2004). The corresponding points then have to be transferred to the range image space (and thus to the 3D object space) having the relative orientation of the RGB and IR cameras from the calibration process. The resulted correspondences in the 3D object space then have to be used for the alignment of the point clouds of different views. The alignment of the point clouds can be improved by an iterative closest point algorithm (ICP) (P.J. Besl and N. MacKay, 1992). To avoid weak alignments, a relatively large overlap between the consecutive views has to be considered. Of course, the accuracy of alignment by this approach depends on the distribution of the SIFT features in the RGB images. Figure 5(b) depicts the aligned point clouds of this example using the presented approach.

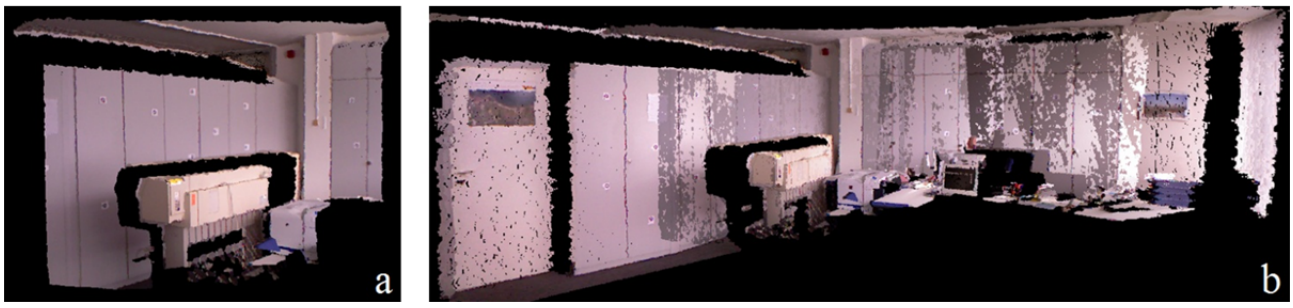

Figure 5: Textured point cloud of a single frame capture (a), point clouds alignment (b)

4. VERY DENSE POINT CLOUD COLLECTION USING A MULTI-CAMERA SYSTEM

4.1. System Design

For the Amsterdam project we decided to develop a multi-camera system and incorporate a dense matching implementation. This enables us to obtain a high resolution point cloud from one single shot. The sensor design was customized to meet certain requirements and restrictions. First of all our assignment was to scan the object with a point sampling distance of 1mm or less and in an accuracy of 1mm or better within twelve days. Furthermore the sensor needed to be small sized and light weighted to be applicable in the given surroundings. The working distance for example was limited to a range of 30cm to 90cm, as the scaffold was very close to the facade and some parts of the object reached into the scaffold area. Another major goal we set for ourselves was to set up a sophisticated system using only low to medium cost components. Also considerations made in the process of project planning and software implementation directly influenced some aspects of the sensor design and vice versa.
The basic concept for the on-site data acquisition was to treat each tympanum as a plane surface in a first step. We planned to move the sensor in a meander shaped trail along the object, capturing imagery (point clouds respectively) as in nadir case airborne photogrammetry. Depending on the

considered hardware components this led to a certain number of necessary sensor positions. Additionally, we estimated that the number of convergent shots needed to capture the rest of the 3D shape of the object in a second step would be three times that high. The total number of stations easily reached several thousands. In order to reduce this number to a minimum it was necessary to cover as much of the object as possible with a single station. This directly implies the usage of short focal lengths.
As for dense matching computation one should take into account, that most approaches are not able to handle large perspectives satisfyingly as they are dependent on high similarity of the image content. This restriction directly leads to the necessity of normal case camera configuration with very high overlap. This constraint is in return highly disadvantageous concerning the achievable measurement accuracy, as it demands short base lengths which again result in bad angular conditions. Our sensor meets this deficiency by increasing the measurement redundancy using four cameras for the matching process. We arrange these cameras in a square with similar viewing directions to achieve a homogeneous measurement quality and a large shared field of view. The scaling of this arrangement mainly depended on the possible working distance, which we assumed to be 70cm in average, and the chosen focal length. The ladder was set to 8mm as this was the shortest focal length available for our cameras.

	Four dense matching cameras	One bundle camera
Type	µEye 2280M	µEye 2250M
Focal length	8mm	4,8mm
Resolution	2448x2048	1600x1200
Pixel pitch	3.45µm	4.4µm

Table 4: Main specifications of the cameras used for our multi-camera sensor

We chose to use industrial cameras as they are very small and light weighted and known to be robust. The main specifications of the cameras can be found in table 4. We defined the overlap of the images to be roughly 90% in both image directions at a working distance of 70cm in order to maintain a sufficient overlap at shorter distances. Basing on this constraint we computed a base length of ~7,5cm. Assuming an image measurement accuracy of 0.3pixels this configuration easily holds the requirements (Table 5).

	30cm distance	70cm	90cm
Ground sampling distance	0.15mm	0.30mm	0.38mm
Triangulation accuracy in viewing direction	0.1mm	0.6mm	1.0mm
Shared field of view of the µEye 2280M	24x19cm	65x53cm	84x68cm
Field of view of the µEye 2250M	35x27cm	80x60cm	100x76cm

Table 5: Pre-estimated specifications of our sensor

Having set up the configuration for single shot data acquisition, the next crucial issue was the registration of the single point clouds. One possibility would have been to directly work on the point cloud data by applying ICP (iterative closest point) algorithms or similar approaches. However, these approaches demand a relatively high overlap between the single point clouds to produce stable results, which in return would have increased the number of sensor stations significantly. To avoid this problem we decided to use photogrammetric bundle adjustment to solve the registration task. Again a high overlap between the images of the different stations is benefitting to the stability of the results. For this purpose we incorporated a fifth camera with even shorter focal

length to the system. The specifications of this camera can also be found in table 4. With an overlap of ca. 60% in both directions of the bundle camera, the number of stations was estimated to roughly 2500 at each tympanum. The stations' positions resulted in a ~30x20cm raster. To further improve the results of the bundle and to solve for global registration we planned a raster of targets on the object. Roughly 50 targets have been placed at each tympanum at a distance of about 50cm vertically and horizontally. These targets have been measured as global coordinates by means of tachymetry and are to be integrated to the registration bundle as control points. Results of this process will be available soon.
All five cameras have been mounted to two custom aluminum adapter plates, which again have been mounted on an aluminum profile. This construction is very compact and provides high stability. For protection of the cameras it was extended by further aluminum profiles surrounding the cameras as a rigid frame (figure 6). The whole construction has a size of ~25x25x15cm and a weight of roughly 2kg (not taking into account the cables).

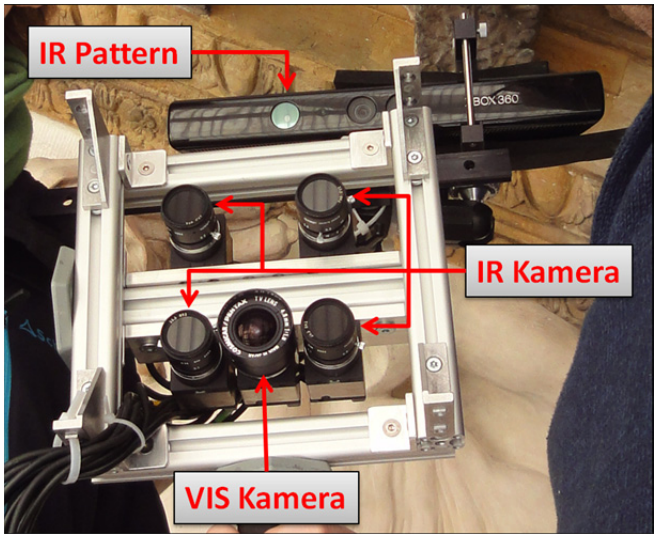

Figure 6: Our Amsterdam sensor. Five cameras rigidly mounted and protected by an aluminum frame. The Kinect device provides additional texture projection.

Figure 7: Image taken by one of the matching cameras. The white speckles covering the lion head's surface are produced by the Kinect's IR laser projector and are not visible to the bundle camera. The image has been slightly enhanced to make the pattern easier to see for the reader.

Enhancement of the object's surface texture was another important issue to be solved in order to generate high quality matching results. Although the freestone surface provided a good texture at large areas there also were areas with homogenous characteristics and also a lot of areas which have been darkened by environmental influences. In such cases, correspondences in the images can either not be found or are ambiguous. In any case, such areas cannot be reconstructed accurately. Thus we decided to use active texture projection in our task. We added a MS Kinect device to our system (see chapter 3), using only its' IR laser projector. Equipping the four matching cameras with 670nm blocking filters made the pattern visible to them while keeping it invisible to the bundle camera. Thus the bundle adjustment process, which incorporates automatic feature point extraction, is not influenced by the pattern moving with the sensor.
At last the system needed to be calibrated accurately. To calibrate the five cameras as one sensor system basically means to define a sensor coordinate system and derive the relative orientations of the cameras with respect to this system. Of course the internal parameters of each camera need to be determined as well. For this purpose we took images (with all cameras simultaneously from each sensor station) of a calibration pattern repeatedly during our work in Amsterdam, mostly following

the known calibration configurations proposed by W. Wester-Ebbinghaus (1983). As a pragmatic computational solution we introduce all images as independent by means of absolute orientation to a classical bundle. We then compute one set of relative orientations per sensor station. The different solutions are then averaged. The standard deviation of the averaged position is ~60µm, for the angles it is ~0,3° (depending on the actual set of calibration data).

	2250M	2280M
C	4.9070mm	8.0286mm
x_0	-0.0547mm	-0.1046mm
y_0	-0.1730mm	-0.0242mm
K1	1.05685e-002	3.64430e-003
K2	1.03265e-004	-8.02600e-006
K3	-2.75286e-007	-7.12524e-007
P1	-2.08425e-004	1.63095e-005
P2	-1.05315e-004	-2.41388e-005

Table 6: Example of derived internal camera parameters. One might notice the strong distortions of the 2250M due to the very short focal length lens.

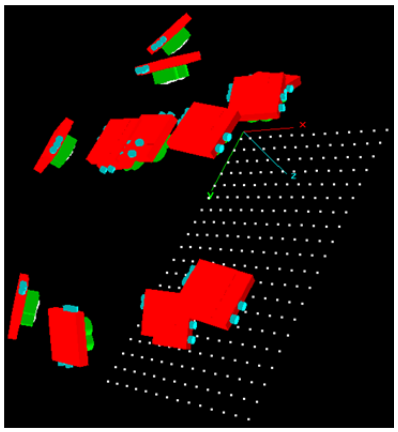

Figure 8: Example of a calibration configuration

However, we assume taking the rigid construction and thus permanent relative orientations directly into account when performing the adjustment benefits to the result's quality. We propose to choose one of the cameras as master camera and define its' coordinate system as the sensor coordinate system. Furthermore we suggest a bundle adjustment approach which introduces the relative orientations of the slave cameras as parameters. This can be achieved by expressing the global exterior orientation of a slave camera as a concatenation of the actual global master orientation and the slave camera's relative orientation. Let us assume a point X, given in world coordinates, is being observed by the master camera. As known, the corresponding image coordinates are found by first transforming the point into the camera coordinate system (X_M) and then projecting it into the image plane. In our case this still holds for the master camera. In case of the slave cameras the procedure changes to first transforming the point into the master camera's coordinate system (global sensor orientation R_0 and T_0), second transforming into the coordinate system of the slave camera (relative slave orientation R_j and T_j) and then projecting to its' image plane. The collinearity equations then change to:

$$x = x_0 - c\frac{r_{11}^j(X_M - X_0^j) + r_{21}^j(Y_M - Y_0^j) + r_{31}^j(Z_M - Z_0^j)}{r_{13}^j(X_M - X_0^j) + r_{23}^j(Y_M - Y_0^j) + r_{33}^j(Z_M - Z_0^j)} + \Delta x$$

and

$$y = y_0 - c\frac{r_{12}^j(X_M - X_0^j) + r_{22}^j(Y_M - Y_0^j) + r_{32}^j(Z_M - Z_0^j)}{r_{13}^j(X_M - X_0^j) + r_{23}^j(Y_M - Y_0^j) + r_{33}^j(Z_M - Z_0^j)} + \Delta y$$

with

$$\begin{pmatrix} X_M \\ Y_M \\ Z_M \end{pmatrix} = R_0^{-1}(\vec{X} - \vec{T}_0)$$

We are currently working on this bundle approach and will present results when presenting the paper within the Photogrammetric Week ’11 program. The presentation will include a comparison with the more pragmatic approach we used so far. During our on-site work in Amsterdam we collected roughly 40GB of image data from ca. 2000 sensor positions, not including calibration data. Depending on the object geometry, a point cloud of 1.000.000 to 2.500.000 points can be produced for each station. The processing of the data is still on-going and will also include a reasonable reduction of the point numbers. Illustrations of first results can be found in the following chapter.

Figure 9:Amsterdam photos (large: camera 1, right from top to bottom: camera 2, 3, 4 and the Master camera 5)

4.2. Dense surface reconstruction from imagery

The multi-view stereo approach employed for our applications uses a dense image matching method providing reliable point correspondences for single stereo pairs. Subsequently, the results of all stereo pairs are merged into one high precision point cloud. However, the key challenge of the dense stereo image matching is the resolution of ambiguities. Therefore, we employ four techniques:

- Epipolar images
- Hierarchical matching
- Patch matching
- Semi-Global Matching optimization

4.2.1. Epipolar images

Using the relative orientation from the calibration of the camera rig the search space for a matching pixel in another image can be reduced to the epipolar line. Therefore, we compute undistorted epipolar images for each stereo pair for the dense matching step, which enables a fast matching along the x axis of the image and the usage of disparity maps which reduces computation duration

and memory requirements. The disparity maps contain the correspondence information and can be stored and accessed efficiently during further processing. Furthermore, the use of epipolar images enables the use of pixel patches for the matching.

4.2.2. Hierarchical matching

Using image pyramids enables narrowing down the search space. Matching images on a low resolution reduces the number of possible correspondences which leads to a lower number of ambiguities and enables large correspondence search ranges over the whole image. Also, the lower resolution ensures the availability of texture required for distinct correspondences. Furthermore, using pixel patches for matching or a global optimization method like Semi Global Matching as described below enables a reliable matching solution. This solution is represented by the disparity map. Mismatches and occlusions are eliminated from this map by validating the disparity value of each pixel using a consistency check and using a speckle filter on the disparity image.
The resulting disparity image for the low resolution is then used in the next pyramid level as initial disparity information. By upscaling the image to twice the width and height and multiplying the disparities by 2 the disparity image can be used to determine a new search range. This disparity range is not determined for the whole image, but instead for each pixel. Therefore, all disparities within a certain mask around the destination pixel (e.g. 5x5 pixels) define the new range.
Consequently, a disparity search range is available for each pixel individually. This is not only beneficial regarding the lower requirements of computation time and memory, but also for the resolution of ambiguities. Within this small search range a matching cost is determined for each possible disparity. The minimum cost represents the final disparity. Sub-pixel accuracy is achieved by estimating a quadratic curve through the minimum costs and the two neighbouring costs and subsequently determining the cost minimum. This cost leads to a new disparity solution for each pixel, which is then again used as initial information for the next pyramid level until the original resolution is reached.

4.2.3. Patch matching

Pixel sets enable a more reliable matching in contrast to the comparison of grey values of single pixels. Ambiguities are resolved by taking into account the pixel neighbourhood and thus, enable the determination of disparities even for small untextured areas. Reliable matching pixel masks can be determined due to the high similarity of the images resulting from the epipolar image projection.
Therefore, we use a Census based matching cost, which is particularly robust against radiometric differences (R. Zabih & J. Woodfill, 1994, H. Hirschmüller & T. Bucher, 2010). In order to use the census based matching costs the image is converted to an image containing the encoded neighbourhood for each pixel. Since only the information of a higher or lower grey values to the center pixel is taken into account this matching cost is highly independent of radiometric differences such as brightness or contrast. Also, not all grey values of the neighbourhood must match to produce a clear result since the magnitude is not important, which is especially beneficially for images with large perspective differences.

4.2.4. Semi Global Matching

The Semi Global Matching method (H. Hirschmüller, 2008) represents a global optimization for stereo matching. The introduction of a global smoothness constraint into the cost function leads to a smoother disparity image and consequently to less visible noise in resulting point clouds. Also, the smoothness constraint leads to a lower sensitivity against noise in the image while small untextured areas can be compensated. This is especially beneficial for matching on small baselines. The

approximation of a global model to 1D paths through the image enables efficient implementations in contrast to other global optimization approaches.
However, the original method consumes a large amount of memory, since cost values must be stored and aggregated for all pixels and subsequently all possible disparities within a certain range. Thus, the number of such of elements is given by W*H*D, where W*H is the image size and D the number of examined disparities within this fixed range. This fixed disparity range is low if the relation between the observed surface undulation and the image scale is low, as occurring for images taken from a far distance to the object. However, for high resolution close range applications the relation between object undulation and distance to the object is very high and thus, requires a large number of disparities to be evaluated. For instance, a 5 Megapixels image would consume for a range of 1000 different disparities 2*2*1000*5000000 Bytes = 12 GB, since two arrays of type unsigned short have to be stored at once. Also, the number of computation steps increase to $16*5*10^9$ for the cost aggregation on 16 paths, as proposed in the original publication about the Semi Global Matching method.
Therefore, we modified the Semi Global Matching method for close range applications by adapting it to the cost storage and evaluation by narrowing down the disparity range for each pixel individually by an image pyramid, as described in the paragraph *Hierarchical Matching*. Instead of assigning and evaluating a cost for the full disparity range only a very small range is examined, which not only leads to very low memory requirements and a small computation time, but also resolves ambiguities and thus, leads to less mismatches.
Even though the Semi Global Matching method provides smoother surfaces and is less sensitive to noise in the image we do not employ this optimization in every case. Since the smoothness constraint works originally only well for planes parallel to the image plane or planes tilted exactly in a direction of one of the paths used during the cost aggregation step, it introduces errors for many close range scenes. Also, it is sensitive against large perspective changes between the imagery as occurring for wide baselines. Furthermore, the optimization requires still much time for high resolution imagery, even though the requirements are lower with the modifications mentioned above.
Thus, we employ the Semi Global Matching optimization method only for the initialization during the hierarchical matching on the first pyramid level for most applications in order to get a good approximate solution. This is not only very fast but also does not require any initial disparity range. On the higher pyramid levels we only use the raw hierarchical matching method. This is sufficient for applications where the ratio between camera base and object distance is sufficient for low noise in the point cloud and where texture is available. The remaining noise is decreased using the redundancy in a multi-view approach.

4.2.5. Multi-view stereo

Multi-view stereo methods use pairwise matching on stereo pairs and fuse the results in a separate step. Many current approaches work in object space which is not suitable for our application since the 3D object space requires a large amount of memory and high computation time for high resolution datasets. Instead, we determine disparity maps for each possible stereo model and use the correspondence information in image space between the stereo models to intersect the optimal object point. This efficient approach scales up to large scenes and enables the determination of high precision point clouds.
Within the implementation stereo matching is performed according to a given connectivity matrix. If no connectivity is available all possible models can be matched due to the fast matching process and subsequently skipped if not enough correspondence information is found. All resulting epipolar and disparity images are stored together with their orientation data. This leads to a certain number of corresponding models for each image appearing as a star-shaped connectivity for each image.

Each of these stars is triangulated by going through each pixel of the base image and all corresponding models. The examined pixel position in the base image is one observation, while each connected model on the star represents another pixel measurement contained in the disparity image of this model.
Together with the orientation data these corresponding pixel measurements are intersected in a common linear triangulation step according to R. Hartley & A. Zisserman, 2008. Remaining outliers and mismatches can be filtered by enforcing each 3D point to be observed at least in 3 images. Thus, all final matches are double-checked using the redundancy. If the residuals exceed a certain accuracy threshold single observations can be eliminated to improve the accuracy and reliability of the resulting point cloud. Using this star approach not only leads to low noise point clouds and range images, but implements a scalable solution due to the low memory requirements and computation time even for large scenes.

4.2.6. Results

(1) As the evaluation of the Amsterdam Tympana is still in progress we want to show first point clouds derived by our dense image matching approach. The implementation during the image data collection was used as in-situ quality control as well, just to check on coarser image pyramid levels the 3D reconstruction by stereo matching in near real-time and to be sure to meet the coverage requirements. As the time constraints for the field work were very strict we had to be sure to collect good radiometric imagery all the time.
Figure 10 represents a point cloud of the upper part of a tympanum. About 58 Million points have been computed from 66 camera stations. It is expected that the object sampling comes close to about 0.2mm with an accuracy in the same range, which is more than sufficient for this application.
(2) In order to demonstrate the applicability of our refined image matching we used the sample data set (C. Strecha et al., 2008) as well (see figure 11). About 33 Mio. points could be computed from 11 images with 6 Megapixels each. Each point was measured in at least 3 images.
(3) The reconstruction of a church in Scotland (St. Andrews) using amateur photos is represented by figure 12. In total, 43 images have been processed delivering a point cloud of about 40 Mio. Points. The resolution of one image was 5 Megapixel.

Figure 10: Extracts from the Amsterdam dataset**.** 58 Mio. points derived directly from the 5 Megapixel bundle camera for about 66 stations. The orientations were derived using Structure from Motion methods. (a) front view onto the point cloud, (b) tilted view.

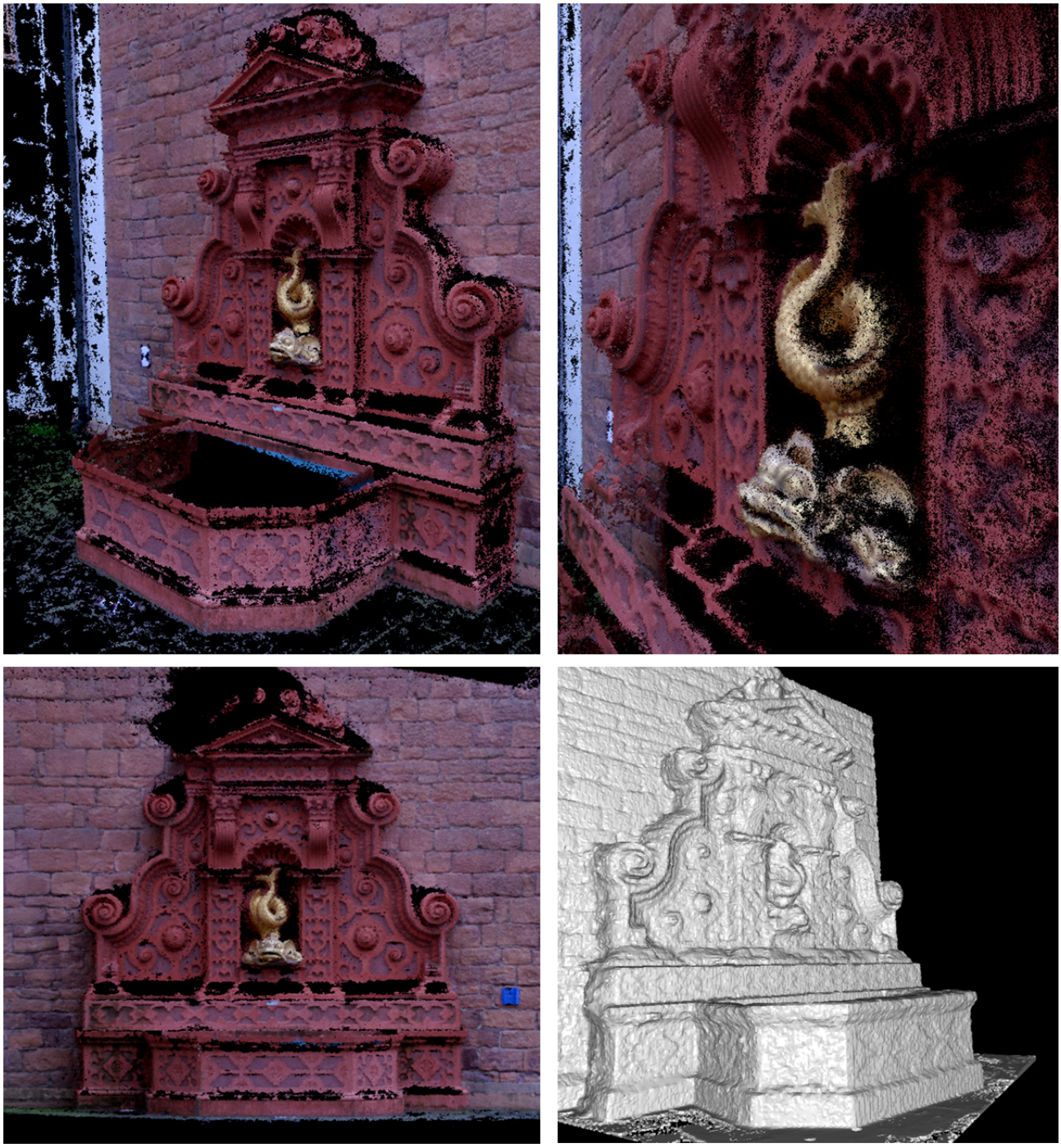

Figure 11: Fountain-P11. 33 Mio. points from 11 images with 6 Megapixels each. Each point was measured in at least 3 images. First images: point cloud, last image: shaded mesh using range image integration and marching cubes. Dataset: MVS evaluation, Strecha et al., 2008.

Figure 12: St. Andrews, Scotland. 40 Mio. points from 43 images with 5 Megapixels resolution each.

5. CONCLUSIONS

The aim of this paper is to demonstrate the usage of low-cost sensor systems, whose data are processed by sophisticated bundle adjustments and refined image matching algorithms. Firstly, we have shown, that smart phones with their integrated sensor systems deliver weak datum parameters to be processed together with their imagery in an extended bundle block adjustment approach. The results are very astonishing and may compete with reconstructions derived by HDS methods.
Secondly, we used the Microsoft Kinect as a sensor system to deliver point clouds for 3D indoors modeling. A simultaneous registration of the MS Kinect imagery with the point cloud was solved just to offer fully textured 3D models. First results are very promising and we will continue this

research line. After studying all the MS Kinect sensors we have used the IR pattern projector for the Amsterdam project.
For the Amsterdam project we were faced with the challenge to deliver very dense point clouds under heavy time constraints for the data collection and to meet the budget requirements. Here we have designed a multi-sensor system with 4 cameras in the corner of a rectangle for several reasons: Firstly to use multi-stereo views, for which up to six stereo pairs could be matched in a pairwise mode. Secondly, we were interested in the potential of multiray photogrammetry using one-shot only to process all rays simultaneously. Thirdly, the success of the data collection had to be guaranteed by ourselves, and therefore high redundancy had first priority.
In order to process the Amsterdam imagery we modified the Semi Global Matching method for close range applications by adapting it to the cost storage and evaluation by narrowing down the disparity range for each pixel individually by an image pyramid, as described in the paragraph *Hierarchical Matching*. Instead of assigning and evaluating a cost for the full disparity range only a very small range is examined, which not only leads to very low memory requirements and a small computation time, but also resolves ambiguities and thus, leads to less mismatches. We have shown, that our implementation is very efficient and have demonstrated its processing capabilities on two further example data sets.
As we just started to demonstrate the potential of low-cost sensor systems and refined image data processing it will be interesting to carry out more projects for the benefit of Cultural Heritage projects, industrial photogrammetry and 3D indoor modelling.

6. REFERENCES

Books and Journals:

Besl, P. J. & MacKay, N. (1992): A method for registration of 3-D shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 14, No. 2, pp. 239-256.

Brown D. C. (1971): Close Range Camera Calibration. Photogrammetric Engineering, Vol. 37, No. 8, pp. 855-866.

El- Rabbany, A. (2002): Introduction to GPS: the Global Positioning System. Artech House Mobile Communication Series, Artech House Inc.

Enge, P. K., Kalafus, R. M. & Raune, M. F. (1988): Differential Operation of the Global Positioning System. IEEE Communications Magazine, Vol. 26, No. 7.

Fritsch, D. & Schaffrin, B. (1981): The Choice of Norm Problem for the Free Net Adjustment with Orientation Parameters. Bolletino Geodesia Science Affini, Vol. 41, pp. 259-282.

Grafarend, E., Kleusberg, A. & Schaffrin, B. (1980): An Introduction to the Variance and Covariance Component Estimation of Helmert Type. Zeitschrift für Vermessung, Vol. 105, No. 4, pp. 161-180.

Haala, N., Fritsch, D., Peter, M. & Khosravani, A. (2011): Pedestrian Navigation and Modeling for Indoor Environments. Proceeding of 7th International Symposium on Mobile Mapping Technology, Cracow, Poland.

Hartley, R. & Zisserman, A. (2008): Multiple View Geometry in Computer Vision. Cambridge University Press, 2nd edition, 6th printing edition.

Hirschmüller, H. (2008): Stereo Processing by Semi-Global Matching and Mutual Information. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30 (2), 328-341.

Khosravani, A. (2010): Digital Preservation of the Hirsau Abbey by means of HDS and Low Cost Close Range Photogrammetry. Master Thesis, Institute for Photogrammetry (IfP), University of Stuttgart.

Kraus, K. (2007): Photogrammetry, Geometry from Images and Laser Scans. de Gruyter, Berlin, 2nd edition edition. 61.

Lemaire, C. (2008): Aspects of the DSM Production with High Resolution Images. IAPRS, XXXVII, Part B4, 1143-1146.

Leick, A. (2004): GPS satellite surveying. 3rd edition, John Wiley & Sons Inc., New York.

Lowe, D. G. (2004): Distinctive Image Features from Scale-invariant Keypoints. International Journal of Computer Vision, Vol. 60, No. 2, pp. 91-110.

Sansonsi, G., Rebeschi, M. & Docchio, F. (2009): State-of-the-Art and Applications of 3D Imaging Sensors in Industry, Cultural Heritage, Medicine, and Criminal Investigation. Sensors 2009, 9, pp. 568-601.

Strecha, C., von Hansen, W., van Gool, L., Fua, P. & Thoennessen, U. (2008): On Benchmarking Camera Calibration and Multi-View Stereo for High Resolution Imagery. CVPR 2008.

Zabih, R. & Woodfill, J. (1994): Non-parametric Local Transforms for Computing Visual Correspondence. Third European Conference on Computer Vision. Stockholm, Sweden.

WWW:

ESA (European Space Agency): http://www.egnos-pro.esa.int/index.html [accessed Mar. 2010].

FAA (Federal Aviation Administration): http://www.faa.gov/about/office_org/headquarters_offices/ato/service_units/techops/navservices/gnss/waas/howitworks/ [accessed Mar. 2010].

Hirschmüller, H. & Bucher, T. (2010): Evaluation of Digital Surface Models by Semi-Global Matching. DGPF-Kameraevaluierungsprojekt: 3-Ländertagung. Wien, July 2010, URL: http://www.ifp.uni-stuttgart.de/dgpf/DKEP-protreff.html.

Wikipedia-Kinect: http://en.wikipedia.org/wiki/Kinect [accessed Apr. 2011].