International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XXXIX-B5, 2012
XXII ISPRS Congress, 25 August – 01 September 2012, Melbourne, Australia

HIGH-RESOLUTION SURFACE RECONSTRUCTION FROM IMAGERY FOR CLOSE RANGE CULTURAL HERITAGE APPLICATIONS

Konrad Wenzel, Mohammed Abdel-Wahab, Alessandro Cefalu, Dieter Fritsch

ifp, Institute for Photogrammetry, University of Stuttgart
Geschwister-Scholl-Straße 24D, 70174 Stuttgart, Germany
{konrad.wenzel, mohammed.othman, alessandro.cefalu, dieter.fritsch}@ifp.uni-stuttgart.de

KEY WORDS: Photogrammetry, Close Range, Cultural Heritage, Multisensor, High Resolution, Imagery, Matching, Point Cloud

ABSTRACT:

The recording of high resolution point clouds with sub-mm resolution is a demanding and cost intensive task, especially with current equipment like handheld laser scanners. We present an image based approached, where techniques of image matching and dense surface reconstruction are combined with a compact and affordable rig of off-the-shelf industry cameras. Such cameras provide high spatial resolution with low radiometric noise, which enables a one-shot solution and thus an efficient data acquisition while satisfying high accuracy requirements. However, the largest drawback of image based solutions is often the acquisition of surfaces with low texture where the image matching process might fail. Thus, an additional structured light projector is employed, represented here by the pseudo-random pattern projector of the *Microsoft Kinect.* Its strong infrared-laser projects speckles of different sizes. By using dense image matching techniques on the acquired images, a 3D point can be derived for almost each pixel. The use of multiple cameras enables the acquisition of a high resolution point cloud with high accuracy for each shot. For the proposed system up to 3.5 Mio. 3D points with sub-mm accuracy can be derived per shot. The registration of multiple shots is performed by *Structure and Motion* reconstruction techniques, where feature points are used to derive the camera positions and rotations automatically without initial information.

1. INTRODUCTION

The recording of high resolution geometry data in cultural heritage applications is a challenging task, especially for large scale objects. Handheld laser scanners often suffer of low acquisition efficiency and of orientation problems, while the costs for such systems are very high. Systems based on structured light projection often suffer from the requirement of a static setup during the acquisition time. Thus, the acquisition efficiency is low as well.

In other applications accuracies around 1-2mm are for most areas sufficient, which can be provided efficiently on large scale by current terrestrial laser scanning systems. But also here, additional recording of high-resolution details is often of interest, which can be done by an image based triangulating system.

The principle of such an image based acquisition system is the stereo image observation. If a point is observed in two 2D images from different views, its 3D coordinate can be reconstructed by intersecting the viewing rays. This method can be extended to multiple images, where the ray intersection not only provides higher accuracy but also high reliability.

In order to acquire images from multiple views, we propose a system where multiple cameras are mounted on a rig with the shape of a square with 7.5 by 7.5cm as described in section 2. For small acquisition distances and high resolution requirements, this shape realizes an optimization between geometric conditions and the concept of image matching for the derivation of 3D data. Due to the compactness of the sensor, the use as a handheld system is possible.

The use of multiple cameras instead of a sequential acquisition of images realizes a one-shot solution. While usual laser scanning systems acquire one 3D point or points along one line at once, image based systems can provide 3D surface information for a whole area due to the matrix structure of the image. This is beneficial for difficult environments where only a small acquisition distance is possible or where vibrations or movements occur. Also, the overall efficiency of acquisition is increased.

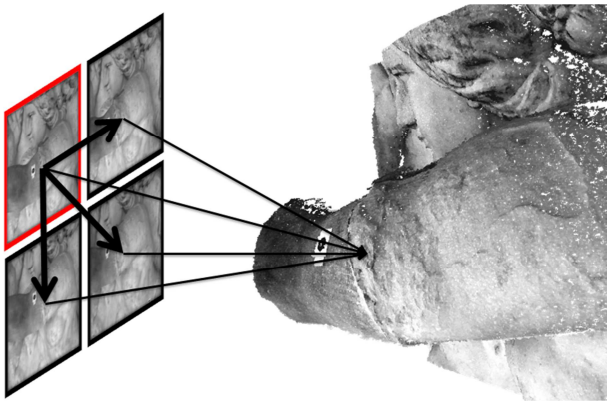

Figure 1: Principle of image based point cloud recording. If images from multiple views are available, corresponding pixels enable the reconstruction of 3D information by intersecting their viewing rays.

The key challenge of subsequently applied image matching methods is to find corresponding points between the images robustly. Therefore, several methods were developed for different applications by the Photogrammetry and Computer Vision community. Feature based image matching methods are extracting distinct features like points or lines, which are identifiable in multiple images. This approach is especially suitable for the reconstruction of 3D information without initial information like applied in *Structure and Motion* reconstruction methods, where camera pose and sparse surface information are automatically derived from images. One popular implementation of this approach is the *Bundler* [Snavely, 2007].

In contrast, dense image matching techniques are using the grey values directly to find a correspondence for almost each pixel of the image such as *PMVS* [Furukawa et al.] or the stereo method *Semi Global Matching* [Hirschmüller, 2008]. The main challenge here is to solve the high ambiguities, since single grey values are not unique in images. Thus, one solution is to apply *Structure and Motion* firstly to derive the exterior orientation. This can be achieved with high accuracy due to the Bundle Block Adjustment as described in section 4. Subsequently, dense image matching methods are applied (section 5), where the epipolar geometry is used to reduce the complexity of the problem and the number of ambiguities.

However, image based acquisition methods usually suffer if no texture on the object to be acquired is available since the image matching fails. Thus, we employ a structured light pattern projector additionally to ensure texture. Therefore, we use the structured light projector of the *Microsoft Kinect* which realizes an encoded pseudo-random pattern emitted by a near-infrared Laser and a diffractive element. The resulting pattern consists of many dots of different size. We use only this projected pattern to support our image matching, while omitting the low-resolution depth information computed by the *Kinect.*

2. SENSOR HARDWARE DESIGN

The optimal design for a multi camera system using image matching techniques is a compromise between the intersection geometry and the performance of the image matching method. The former requires a large angle for high accuracy while the latter performs best at small angles, where the image similarity is high. Thus, if the angle is small the completeness of the resulting point cloud is higher since the matching performs better and the amount of occluded areas in the overlapping imagery is reduced.

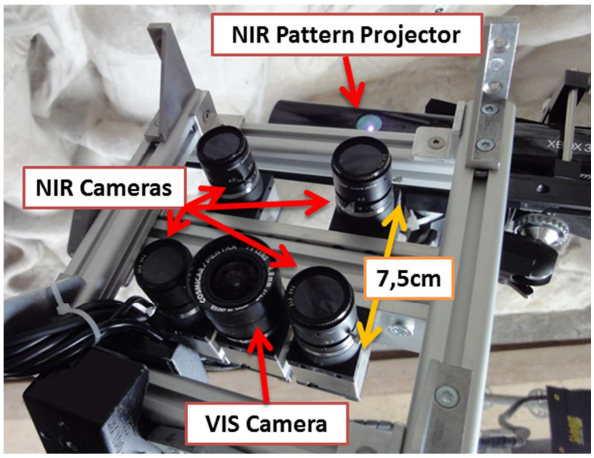

Figure 2: Sensor design overview

The four cameras used for the dense image matching are arranged in a square with the size of 7.5 by 7.5cm as shown in figure 2, optimized for a good matching performance and high accuracy at a short acquisition distance of about 70cm. They are equipped with a 5 Megapixels CCD and a lens with 8mm focal length, which leads to a footprint of 65cm by 53cm at 70cm distance and an estimated depth precision of about 0.84mm, as shown in table 1.

Since the measurements are derived based on the triangulation principle, the depth precision varies with the distance to the object. The relation between the distance to the object H, base B and focal length f and precision in image space σ_d is given by [Kraus 2007] as follows:

$$-H = \frac{fB}{x_p - x_q} = \frac{fB}{d}$$

Thus, the precision of depth can be approximately estimated in respect to the precision in image space σ_d while neglecting the errors of the exterior orientation as follows:

$$\sigma_H = \frac{\partial H}{\partial d} = \frac{H^2}{fB}\sigma_d$$

Thus, for the proposed system the depth precision can be estimated as shown in table 1, where a precision of the dense matching method in image space of 0.3 pixels [Rothermel et al., 2011] is assumed.

Distance	**[cm]**	**50**	**70**	**100**	**120**
Precision	**[mm]**	0.43	0.84	1.73	2.48
Resolution	**[px/mm²]**	21.5	11.0	5.4	3.7
Footprint	**[cm]**	45*36	65*53	98*81	119*98

Table 1: Theoretical approximate precision in object space, resolution and footprint size for different acquisition distances

As mentioned before, a *Microsoft Kinect* is used to project artificial texture onto the object. Only the pseudo-random pattern projector, powered by a near-infrared Laser is used to support the image matching procedure. Since the projector is compact, it can be mounted on the camera rig as well and consequently be moved with the sensor. This enables data acquisition even for low light conditions or for objects totally free of texture.

In order to make the pattern visible to the four matching cameras, a 670nm daylight blocking filter is mounted to the lenses. This wavelength also enables the hybrid acquisition of the materials texture and the artificial texture together since the filter is still transparent to a part of the daylight. This enables optimal matching conditions by providing distinct grey values for each pixel in arbitrary environments.

However, the artificial pattern is moving with the sensor and consequently would disturb a registration of multiple shots by feature points in the image. Thus, a fifth 2 Megapixels camera with a very short focal length of 4.7mm was mounted on the rig. It has a very large field of view with 105cm width and 78cm height at 70cm distance. Consequently, less overlap is required for the registration of multiple shots by image features. Since it has no daylight blocking filter but only an infrared blocking filter, the speckle pattern is not visible in the images and does not disturb the feature point extraction and matching.

Consequently, the actual data acquisition can be performed sequentially, where 5 images are acquired at each station. With an overlap of about 70% between the images acquired by the wide angle camera of subsequent shots, the dense point clouds can be connected. Finally, this leads to one dense point cloud.

3. SOFTWARE PIPELINE OVERVIEW

In order to derive dense 3D point clouds from the acquired imagery the interior and exterior orientation parameters have to be determined as described in section 4. Subsequently, the dense image matching is performed on stereo pairs (5.2), followed by a multi-baseline triangulation step (5.3) to derive dense 3D point clouds. Finally, post processing steps are applied containing filtering methods in object space to clean up the cloud from remaining outliers.

4. ORIENTATION DETERMINATION

The **interior orientation**, consisting of the camera parameters like focal length and lens distortion, should be determined by a calibration to meet high accuracy requirements. This calibration can be performed using a standard method employing a calibration pattern. The **relative orientation** between the cameras within the rig can be estimated by the bundle adjustment of the pattern calibration as well.

The **exterior orientation** of the whole camera rig, which corresponds to the registration of multiple shots, is realized by a *Structure from Motion* reconstruction method as described by Abdel-Wahab et al., 2011. In the images of the wide angle camera which is not observing the *Kinect* pattern, *SIFT* [Lowe, 2004] feature points are extracted. Subsequently, the descriptors of these points are matched to the ones of other images, which enables the computation of the relative orientation. By performing this step sequentially for each neighboring image pair followed by a bundle adjustment after each extension, the whole block can be connected.

Finally, the exterior orientation for all cameras can be extracted together with a sparse point cloud representing the triangulated feature points. Despite an approximate value for the focal length no initial information is required for the whole process. However, the key challenge is the high complexity of the image network especially for large scenes. Within the work of [Abdel-Wahab et al., 2011] an initial network analysis is performed to overcome this problem.

The **scale** can be introduced to the bundle by a scale bar or preferably by a set of ground control points. Furthermore, using ground control points supports the bundle adjustment and enables georeferencing.

If the calibration of the relative orientation within the rig and the interior orientation shall not be repeated or automation is required due to slight changes during a long term use, it can also be derived automatically. Therefore, a registration by feature points followed by a bundle adjustment can be used to re-estimate the relative and interior orientation between the cameras, if a scene with sufficient texture is available since the *Kinect* can't be used. However, an initial calibration is beneficial to remove the lens distortion, which increases the number of successfully matched feature points and thus supports the bundle adjustment.

5. DENSE IMAGE MATCHING

5.1 Strategy

The dense image matching step is used to derive dense 3D point clouds from the acquired imagery. This is achieved by the determination of pixel-to–pixel correspondences between the images of the camera rig. Therefore, a base image is selected similar to the red framed one shown in figure 1. Subsequently, for each pixel of this base image a correspondence to all the other 3 images is determined. The resulting 4 viewing rays are intersected in space and lead to one 3D point in a triangulation step. Consequently, a 3D point is derived for almost each pixel in the overlapping part of the base image. Finally, filters are applied to the resulting point cloud within a post processing step. This whole process can be applied on arbitrary image acquisition configurations.

5.2 Dense stereo method

The dense stereo method determines the correspondences between each stereo image pair. Therefore, we developed a hierarchical matching method using a modified approach of *Semi Global Matching* [Hirschmüller, 2008]. The modifications are applied to increase the efficiency while enabling the processing of high resolution imagery with very large depth variations as occurring for this application.

The *Semi Global Matching* stereo image matching algorithm approximates a global smoothness over the depth solution, in order to get a solution consistent with the global neighborhood. The resulting smooth surfaces have low noise while ambiguities can be resolved robustly. Due to the approximation of the global smoothness constraint as 1D paths through the image, the algorithm can be implemented efficiently also for high resolution imagery. Furthermore, epipolar images can be used to reduce the complexity since the matching is only performed along the x-axis.

In the original implementation of *Semi Global Matching* so called *matching costs* are assigned in a cube shaped data structure. For each pixel and each possible correspondence the similarity of the two grey values is stored as matching cost. The cost is low if the similarity is high and vice versa. We use the *Census* matching cost [Zabih & Woodfill, 1994], which is robust since it has low ambiguities due to the mask used around the evaluated pixel. Also, it is highly insensitive against radiometric changes.

The possible correspondences of a pixel are described as disparities along the epipolar line. This range is fixed over the whole image. However, with an increased resolution of the image and a close acquisition distance the depth and thereby the disparity variations become very large. Thus, the evaluation becomes complex and the number of ambiguities increased. Also, the cost array can't be stored in the memory of current workstations.

In order to solve this problem we developed a hierarchical approach, where the images are analyzed on several resolution levels. From the lowest resolution level to the higher ones the disparity range is narrowed down and assigned for each pixel individually. This dynamic approach reduces the number of ambiguities, the memory consumption and the processing time.

Within the implementation the number of pyramid levels is selected according to the resolution of the image. On the highest pyramid level the dense image matching is performed on a relatively large disparity range covering almost the whole image. Thus, no initial depth information is required. However, due to the low resolution the actual amount of evaluated disparity samples is very small. The image matching is performed in both directions in order to apply a consistency check for the elimination of outliers or occluded areas. The resulting disparity images are passed to the next pyramid level. After resizing the disparity image and multiplying it by 2, it is used for the determination of the disparity range which is done for each pixel individually. Within a certain mask (e.g. 5 by 5 pixels) the minimum and maximum of the disparity values is determined and extended by a buffer. This defines the new disparity range for the subsequent matching step on the current pyramid level. Only this range is evaluated within a dynamic data structure supporting different depth ranges for each pixel.

The filtering after each pyramid level is important to remove outliers, which would disturb the disparity range. Beside the left-right consistency check, a speckle filter is applied to remove disparity speckles with an area smaller than 100 pixels.

5.3 Multi-Stereo Point Triangulation

Each disparity map derived from the previous step describes the pixel-to-pixel correspondences between a stereo pair. In order to retrieve the 3D point coordinates for each of these correspondences, the two viewing rays could be reconstructed for each correspondence by using the exterior orientation. Thus, the intersection could be performed for each disparity available in the map which leads to one dense 3D point cloud.

However, accuracy and reliability of the point cloud can be increased significantly if more than two rays are available. In order to reconstruct more than two rays for an observed 3D point the correspondences between the different stereo pairs have to be evaluated.

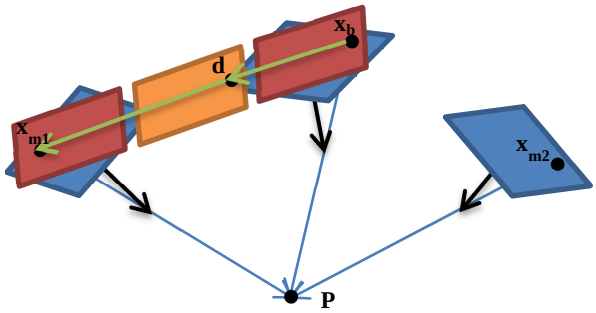

Figure 3: Multi-Baseline Triangulation Principle.
One 3D point observed in 3 images (blue). Within a stereo model the correspondence is given by the disparity image (yellow), which is related to the epipolar images of the model (red).

Figure 3 shows the applied Multi-Baseline Triangulation principle. Three images (blue) observe a common point. In order to compute such a 3D point for each pixel in the base image (middle), this base image is matched to the left image and the right image. Therefore, epipolar images (red) are used within each stereo model. In order to determine the correspondences to a pixel in the base image $\mathbf{x_b}$, the epipolar homography matrix is used to transform the pixel coordinate to the coordinate in the epipolar image of the model. The disparity image (yellow) stores the disparity **d** which leads to the corresponding pixel in the match image $\mathbf{x_m}$ of the model. This correspondence is estimated for each available model (here 2).

With the exterior orientation of the base image and the base image measurement $\mathbf{x_b}$ as well as the exterior orientation of the epipolar match image and the corresponding measurement $\mathbf{x_m}$ the 3D point **P** can be triangulated. Therefore, a linear triangulation method such as proposed by [Hartley et al., 2008] can be used. By eliminating reprojection errors greater than 1 pixel outliers can be rejected, which increases the reliability of the point cloud. Also, intersection angles smaller than 3 degrees are suggested to be eliminated, since the glancing intersection leads to noisy results.

For the presented sensor system the upper left camera was defined as base camera. Consequently, 3 stereo models are available for image matching and triangulation.

6. EVALUATION

The accuracy for the proposed system is optimal if texture is available. Especially natural materials like stone or wood provide good texture so that each pixel has a distinct grey (color) value. This is important since the image matching accuracy is reduced if two neighboring pixels have the same value. For single pixels or small patches this is usually no problem since the smoothness constraint of the *Semi Global Matching* algorithm interpolates a suitable depth using the textured neighborhood.

For classic cultural-heritage data recording sufficient texture is usually provided. For example, already the grain of marble like the one in figure 4 is suitable since digital cameras provide high radiometric sensitivity. Thus, all materials which are not reflecting light too much are recordable. However, in urban or industrial environments texture-free areas occur for clean materials such as white walls without surface structure or plain plastic.

Figure 4: Data recording example (project: see section 7).
Upper left: Color image of the scene acquired by DSLR.
Upper right: grey image acquired by the sensor. The white rectangle depicts the magnified area of the lower left image, showing the texture with the *Kinect* speckle pattern. Lower right: computed depth image.

Therefore, the proposed system is equipped with an artificial texture projector, represented by the *Microsoft Kinect* as described in section 2. It projects an equally distributed speckle pattern with dots of different size and brightness as shown in figure 4 and 5. This enables data recording on arbitrary surfaces with high depth resolution. Again, the only limitation are reflecting materials which do not provide a visible texture on the actual surface. Slightly reflecting materials can be recorded as long as the pattern is visible to the camera. Furthermore, the strong laser of the *Kinect* enables recording also on dark materials, since sufficient light is reflected to be resolved by the sensitive sensor. However, during classical cultural heritage data recording the pattern is only supporting while the texture of the material enables high resolution depth estimation already.

For objects with texture, accuracy in depth of about 0.8mm is reached for an acquisition distance of 70cm. However, the performance of the system for critical conditions where no

texture is available shall be investigated. Consequently, texture is only provided by the speckle pattern of the *Kinect*.

The evaluation setup, as shown in figure 5, contains several objects like spheres and planes of different material. Images were acquired and processed for different settings to investigate the impact of parameters like exposure time, external illumination, number of cameras, base length or acquisition distance. The most challenging object is a white plastic sphere without own texture. The projected speckle pattern is not providing texture for the whole area, but only by the particular dots of different brightness. Thus, small untextured areas occur which lead to an increased noise in the point cloud.

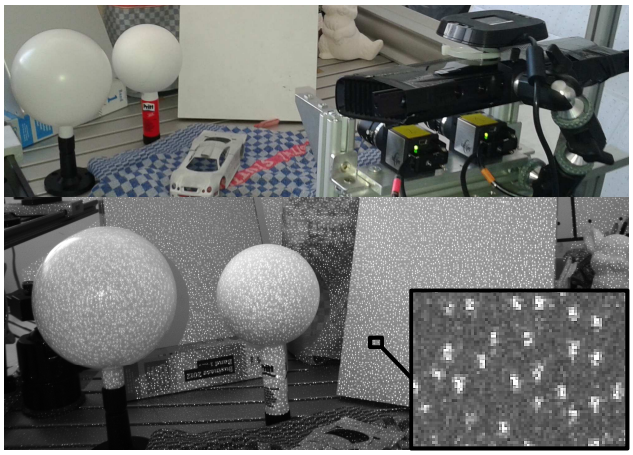

Figure 5: Evaluation setup (upper) and camera images (lower) with *Kinect* speckle pattern. Lower right image: magnified image detail.

In order to determine the accuracy of the system in object space, images were acquired for the sphere (r=15cm) for different illumination conditions at a distance of about 70cm. This is also challenging for the image matching method, since the smoothness constraint of the *Semi Global Matching* algorithm (section 5.2) works only for constant neighboring disparities. In contrast, the sphere provides continuous depth changes and thus enables the evaluation of the most challenging condition. The radius of the sphere was determined under optimal conditions using a *Gauß-Helmert* parameter estimation. Subsequently, the radial residuals to the actual sphere surface can be determined.

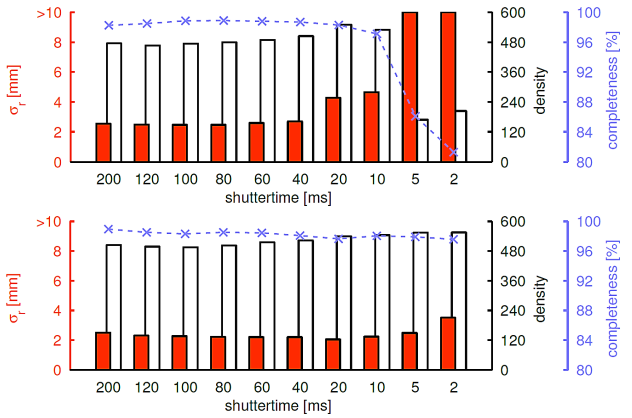

Figure 6: Accuracy in object space for a reference sphere without texture. Only the *Kinect* provides texture which leads to noisier, but still suitable results. The x-axis depicts the shuttertime, the left y-axis the standard deviation of the radial residuals. The upper figure shows the evaluation for images with 8 Bit depth, the lower one for 12 Bit.

Figure 6 shows the result for different exposure times if only the *Kinect* is used as light source. As can be seen, the behavior of the accuracy is stable if the exposure time is sufficiently long. Especially if a high radiometric depth of 12 Bit is used, also very short exposure times are possible. For a 12 Bit depth the camera can discretize 4096 different grey values instead of 255 only and thus use the full radiometric capabilities of the sensor. This is particularly beneficial for low-light conditions and thus is suggested if the system shall be used for hand-held applications. The mean standard deviation of the radial residuals to the sphere amounts to 2.2mm. However, significantly better accuracies can be reached if the object surface provides texture on its own since no texture-free areas occur as for the speckle pattern.

7. CULTURAL HERITAGE APLLICATION

The proposed sensor system was used for a large scale cultural heritage data recording project at the Royal Palace in Amsterdam. Two tympanums covering an area of about 125m² were recorded with sub-mm resolution and accuracy. The tympanums contained reliefs with a complex 3D surface as shown in figure 7. A first comprehensive report about this project is given by Fritsch et al., 2011.

Figure 7: Upper left: East tympanum at the palace of Amsterdam from distance. Upper right: detail - the relief contains whole statues (size of visible target: 4cm by 3cm). Lower: sensor during data acquisition.

In order to acquire a dense point cloud for each tympanum the designed sensor was used at the scaffolding in combination with several tripods and stands. Firstly, a nadir acquisition was performed in a raster shape to cover both triangle shaped areas with the size of 25m by 5m each with connected point clouds. Secondly, angled shots were acquired to record details and to resolve occlusions. The compact size of the system was particularly beneficial for this task.

The fifth camera with the very large field of view was not used for the actual point cloud acquisition but for the registration of point clouds only (see also section 4). Feature points were used within a *Structure from Motion* method to determine the exterior orientation of the stations without initial information. Within 9.5 days about 2,000 stations were acquired leading to a total amount of about 10,000 images. Ground control points measured by tachymetry provided the transformation to the global coordinate system. About 2 billion 3D points were computed for both tympana.

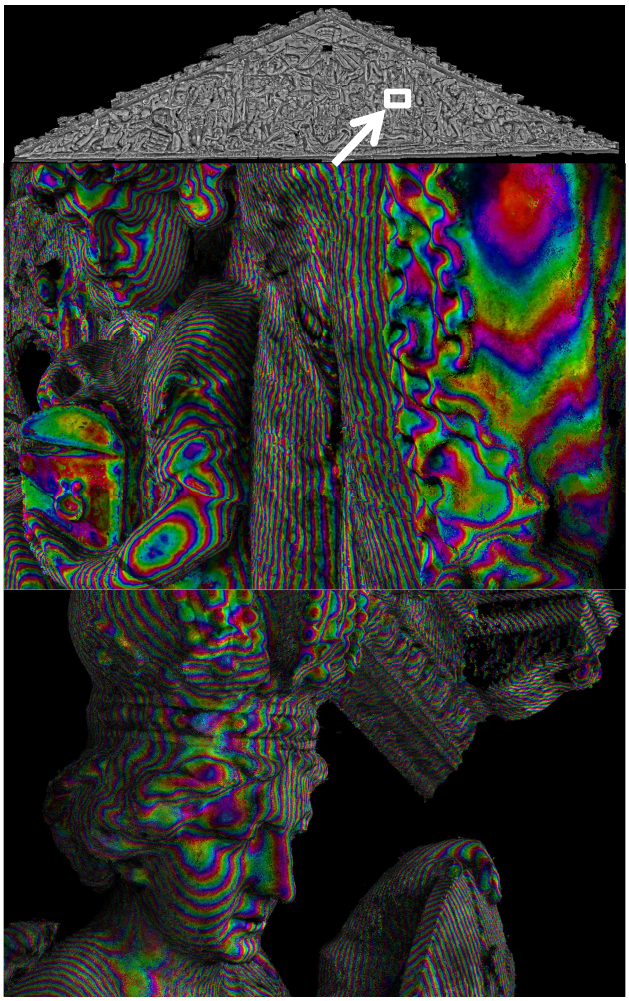

Figure 8: Point cloud of the tympanum at the west façade with about 1.1 billion points and a detail. Third image: detail from the east tympanum.

8. CONCLUSION AND OUTLOOK

Image based data recording provides an effective solution especially for cultural heritage applications with high accuracy requirements. Cameras can be used for hand-held solutions which enable the efficient acquisition of complex details with high resolution while occlusions can be resolved by multiple viewpoints. At the same time, cameras are affordable off-the-shelf products, while common systems based on laser scanning are expansive and do not provide the same level of flexibility.

The proposed sensor system is a realization of this concept and proved that even a recording on large scale is possible. Dense image matching methods provide 3D information for each pixel in the overlapping areas which leads to a point cloud of up to 3.5 Mio. points per shot. The mounted *Microsoft Kinect* ensures the availability of texture, which enables also low light applications or the recording of surfaces without texture. However, higher accuracies can be reached if the object has texture on its own, which is the case for most cultural heritage data acquisition tasks.

Within future work, the investigation on alternative ways for the real time derivation of the exterior orientation can enable an efficient acquisition work flow with preview and guidance functionalities. Therefore, *Simultaneous Location and Mapping* (SLAM) methods can be employed or a real time registration point cloud in object space by *Iterative Closest Point (ICP)*. Also the low resolution but real-time depth information from the *Kinect* might be beneficial for these tasks.

9. ACKNOWLEDGEMENTS

We would like to thank Matthias Küver for providing the results of his extensive accuracy study on the proposed sensor system. Furthermore, we would like to express our gratitude to Erwin Christofori, Jörg Bierwagen and the team from the surveying company Christofori and Partner for the pleasant cooperation within the project in Amsterdam. Thanks to Thomas Zwölfer and Patrick Tutzauer for their assistance during the data acquisition.

10. REFERENCES

Abdel-Wahab, M., Wenzel, K., Fritsch D. 2011. Reconstruction of Orientation and Geometry from Large Unordered Image Datasets for Low Cost Applications. Low-Cost 3D (LC3D) workshop, December 2011

Fritsch, D., Khosravani, A., Cefalu, A., Wenzel, K, 2011. Multi-Sensors and Multiray Reconstruction for Digital Preservation. In: Photogrammetric Week ’11 (Ed. D. Fritsch), Wichmann, pp. 305-324.

Furukawa, Y., Ponce J., 2007. Accurate, Dense, and Robust Multi-View Stereopsis. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, July 2007.

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), pp. 328–341.

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

Lowe, D., 2004. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision, Vol. 60, pp. 91-110

Rothermel M., Haala, N., 2011. Potential of Dense Matching for the Generation of High Quality Digital Elevation Models. In ISPRS Proceedings XXXVII 4-W19

Snavely, N., Seitz, S. and Szeliski, R., 2007. Modeling the world from internet photo collections. International Journal of Computer Vision.

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