SURE: PHOTOGRAMMETRIC SURFACE RECONSTRUCTION FROM IMAGERY

Mathias Rothermel, Konrad Wenzel, Dieter Fritsch, Norbert Haala

Institute of Photogrammetry
University of Stuttgart
Stuttgart, Germany
Geschwister-Scholl-Str. 24D, 70174 Stuttgart, Germany
prename.surname@ifp.uni-stuttgart.de
http://www.ifp.uni-stuttgart.de

KEY WORDS: Surface, Reconstruction, Software, Aerial, Close Range

ABSTRACT:

This article presents an approach of a multi-view stereo (MVS) method for the generation of dense and precise 3D point clouds. It is based on the Semi-Global Matching (SGM) method followed by a fusion step in which the redundant depth estimations across single stereo models are merged. We present a hierarchical coarse-to-fine solution for the SGM method in which matching results of low resolution pyramids are used to limit disparity search ranges for high resolution pyramids. By means of large format aerial and close range imagery we show that memory demands as well as processing times can be significantly decreased whereas the quality of resulting disparities estimations is maintained. By merging redundant disparity estimations of multiple stereo models the precision and robustness of the generated point clouds can be increased. Based on basic principles of epipolar geometry we present a time efficient algorithm for outlier detection and object point triangulation minimizing the reprojection error. Thereby the geometric configuration of adjacent cameras is taken into account. An implementation of the algorithm called SURE as well the library interface libTSgm of the presented algorithm is publicly available at: `http://www.ifp.uni-stuttgart.de/publications/software/`.

1 INTRODUCTION

3D reconstruction of real world objects using imagery has been a vivid research area for decades in computer vision as well as photogrammetric community. Nowadays applications range from the generation of high resolution surface models using large frame aerial or UAV imagery, object modeling in the film and game industry, cultural heritage documentation, surveying for quality control, up to driver assistance systems claiming for real time performance. Premising good geometric configuration of views, sufficient accuracy of interior and exterior orientations and good radiometric quality of imagery, state-of-the-art MVS solutions reconstruct depth estimations for nearly each image pixel providing precisions in the sub-pixel range (Remondino and Zhang, 2006) (Haala and Rothermel, 2012a). Even for radiometric low quality imagery providing extreme diversity in image content as available from online photo communities models of compelling detail and size can be reconstructed (Merrell et al., 2007) (Goesele, 2007). Despite a huge number of MVS solutions have been published, only few number are publicly available. Foremost PMVS, an implementation of a patched based MVS (Furukawa and Ponce, 2010) recently gained a lot of attention. This surface growing algorithm initialize surface patches and the respective orientations based on salient feature points. In an expansion step the surfaces around these patches are reconstructed. In this implementation all images are used simultaneously which implies large memory demands. However, this issue can be overcome by clustering the input images and then reconstructing sub spaces of the scene as proposed in (Furukawa et al., 2010). As a second example of freely available software the MicMac (Pierrot-deseilligny and Paparoditis, 2006) package implements a coarse-to fine modification of the maximum flow matching algorithm proposed by (Roy and Cox, 1998). Thereby global cost function for the multi-view correspondence problem is formulated as maximum flow problem. The minimal cut then represents the surface minimizing the global cost. In contrast, according to the taxonomy of (Seitz et al., 2006), the MVS described in this article is classified as image space method. Within our approach a reference image is matched to a set of adjacent images using a SGM-based stereo method. For each pair a disparity map is computed. Afterwards all disparity maps sharing the same reference view are merged. One of the advantages of this class of algorithms is that it scales well to large datasets. However, in order to capitalize redundancy across single stereo pairs a proper and time efficient fusion method has to be provided. Note that the presented MVS does not implement an multi-photo consistency measure (A. Gruen, 1988) (Okutomi and Kanade, 1993), instead photo consistency measures are based on single stereo pairs and geometric consistency constraints are imposed at the fusion stage.

Dense stereo matching within our implementation is based on the SGM algorithm (Hirschmüller, 2008). Due to its dense reconstructions preserving disparity discontinuities, high robustness regarding parametrization and real-time capability, SGM often is the technique of choice for real world applications. However, memory demands are extensive since photo-consistency information of all pixels and their sets of potential correspondences have to be kept in memory for the subsequent semi-global optimization. For the optimization step itself a second buffer for aggregated costs of the same size is required. For large format frames and scenes possessing large variances in depth, the method demands for drastic memory consumption. This problem was recently addressed in (Hirschmüller et al., 2012) where the aggregated costs are stored only for the eight most probable candidates. Although it is stated that results are of same quality than the classical approach, the method comes at the price of increased processing times (theoretically factor 1.5). As memory demands and processing times for the optimization steps scale with the number of potential correspondences to be evaluated, more efficient SGM modifications operating on limited disparity search ranges were proposed. The first coarse-to-fine modification for SGM was presented in (Gehrig et al., 2009). In this work disparity priors from low resolution imagery are used to derive a region of interest (ROI) representing far objects. This ROI is matched on full resolution using a limited but constant disparity search

range. Then disparity priors and ROI results are merged. This is based on the idea that disparities of image parts representing close objects were estimated sufficiently accurate in low resolution matching cycles. In (Hermann and Klette, 2012) disparity priors are used to initialize disparity search ranges of individual pixels for matching subsequent stereo pairs. Although a speed up of the matching process is attained, the algorithm still operates on constant disparity range buffers and does not limit the memory requirements at all. Our stereo approach is most similar to (Hermann and Klette, 2013) and (Wenzel et al., 2011) where disparity priors are used to derive search ranges for each pixel individually and moreover the size of all required buffers are dynamically adapted. In section 3 we show that this leads to significant reduction of memory and time requirements. A central problem in hierarchical approaches and search space reduction is to estimate ranges which allow to recover from erroneous priors and at the same time minimize memory consumption and computational complexity. In (Hermann and Klette, 2013) the concept of semi-global distance maps is introduced for search space determination. In contrast we apply a simple window based analysis locking valid disparities for the next level and reinitializing disparities and the corresponding ranges if matching failed. Details of the stereo approach are discussed in section 2.2.2.

Within image space MVS methods disparity estimations of single stereo models are typically merged in order to increase the reliability and precision of the final depth maps or point clouds. The methods in (Goesele et al., 2006) utilize a volumetric approach for depth map fusion based on (Curless and Levoy, 1996). Thereby depth maps are used to construct a signed distance field from which a isosurface is extracted. This isosurface can be efficiently converted to a triangular mesh using for example the marching cube algorithm (Lorensen and Cline, 1987). In (Fuhrmann and Goesele, 2011) an algorithm capable of fusing depth maps of image sets possessing large variances in image scale was proposed. It is based on a hierarchical signed distance field enabling the representation of surfaces with varying detail. The mesh-based approach proposed in (Turk and Levoy, 1994) removes or fuses triangles of overlapping regions across two depth maps. Then resultant sub-meshes are glued together. (Merrell et al., 2007) fuse a large number of limited quality stereo depth maps by claiming geometric consistency incorporating confidence measures available from image matching. In a following step redundant depth estimations of fused maps are merged to one vertex and the final mesh is generated using a quad tree method. Our approach is most similar to (Koch et al., 1998) which introduced correspondence linking technique for disparity map fusion in sequence of images. Disparity maps are generated for each reference view and its two adjacent views. Using homographies redundant measurements are linked across multiple views in the sequence. Outlier detection and inlier fusion for a set of redundant disparities is performed using a Kalman filter. Because of the requirement to deal with unstructured image collections and to incorporate a larger number of stereo models, within our approach a reference view is matched against multiple adjacent views. Within this cluster of views redundant correspondences are linked and checked for geometric consistency. Consistent measurements are then fused minimizing the reprojection error. Using basic principles of epipolar geometry we express the problems of consistency check and triangulation in dependence of the depth only, which enables efficient computation. The algorithm is explained in detail in section 2.3. The result is one accurate depth image or point cloud per reference view. Although generated point clouds are of good quality redundancy could be further exploited by depth image integration.

2 ALGORITHMS IN SURE

Given a set of oriented input images the SURE-algorithmic extracts 3D point representing the scenes surface. The implemented tool chain is split-up into four main modules as displayed in figure (1). In this section first a general outline of the tool chain is given. Next, the three main software modules are described in detail. A preprocessing module performs a network analysis and selection of suitable image pairs for the reconstruction process and is only shortly covered in this article.
Within a first main processing step epipolar images are generated for each stereo pair. Within the second step dense matching is carried out on the generated epipolar images. Within this step disparities/parallaxes across stereo pairs are calculated. Thereby the SGM method was modified in order to enable a time and memory efficient processing. Within our tool chain an image $\mathbf{I}_b$, in the following referred as base image, is matched against a certain number N of proximate (match) images resulting in a set of stereo models $\mathbf{M}_{n=1,\ldots,N}$. For datasets possessing high overlaps of incorporated imagery, within these stereo models depth information of the surface is estimated redundantly. In the third module this redundancy is exploited to eliminate blunders and increase the accuracy of depth measurements. Thereby only depth maps of stereo models sharing the same base image, $\mathbf{I}_b$,$\mathbf{I}_{m,i=1,\ldots,n}$ are fused. The result is a depth image (or point cloud) with respect to the base image $\mathbf{I}_b$.

The information, which stereo models should be incorporated into the reconstruction process, is defined by connectivity matrices. These are stored as ASCII files and passed to the single modules. For small datasets and structured image configurations the stereo models to be incorporated might be obvious and connectivity matrices can be specified manually. However, for large and unstructured image collections this task is not trivial. Therefore, a method for the initialization of the image network has been developed, which derives and filters the connectivity information using the exterior orientation of the images. For some SFM/BA packages the connectivity information is already available. If necessary it can be thinned out by thresholding or based on baselines limitation or analysis of intersection angles of the principle camera rays. If this connectivity information is not available a 3D reconstruction for the dataset using low resolution images is carried out. This low resolution imagery enables fast processing. Then, based on the generated (sparse) 3D surface the actual overlaps, scale differences and angles across different stereo pairs can be derived and pairs suitable for processing can be automatically determined.

2.1 Rectification Module

Within the image rectification module epipolar images for the matching process are generated. Let $\mathbf{I}_b$ and $\mathbf{I}_m$ be a pair of images to be rectified and $\mathbf{I}_b^r$ and $\mathbf{I}_m^r$ the resultant epipolar images. $\mathbf{I}_b^r$ and $\mathbf{I}_m^r$ are virtual images providing the same optical centers as original $\mathbf{I}_b$ and $\mathbf{I}_m$ but posses updated rotations and internal parameters. Rectified and original orientations then define the two 3×3 matrices $\mathbf{H}_b$, $\mathbf{H}_m$ relating the homogeneous image coordinates $\mathbf{x}_b$, $\mathbf{x}_m$ in the original images and $\mathbf{x}_b^r$, $\mathbf{x}_m^r$ in rectified images according to

$$\begin{aligned} \mathbf{x}_b^r &= \mathbf{H}_b \mathbf{x}_b \\ \mathbf{x}_m^r &= \mathbf{H}_m \mathbf{x}_m. \end{aligned} \tag{1}$$

The inverse mapping can be simply calculated as

$$\begin{aligned} \mathbf{x}_b &= \mathbf{H}_b^{-1} \mathbf{x}_b^r \\ \mathbf{x}_m &= \mathbf{H}_m^{-1} \mathbf{x}_m^r. \end{aligned} \tag{2}$$

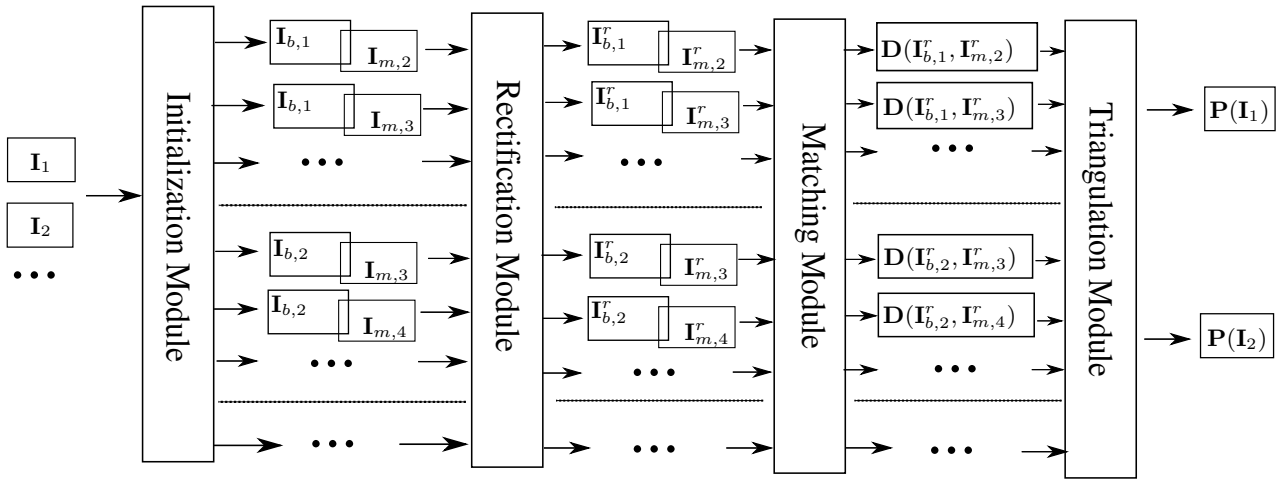


Figure 1: Flow chart of the implemented algorithm. In the initialization module stereo pairs to be incorporated into the reconstruction process are chosen. Selected stereo pairs then are rectified and matched. Eventually disparity maps which correspond to the same base image are fused and resulting depth image or point cloud are calculated.

For the central task of finding the homographies $\mathbf{H}_b$ and $\mathbf{H}_m$ two algorithms are implemented. Using the approach of (Fusiello et al., 2000) the rotation of the image planes $(\mathbf{I}_b, \mathbf{I}_b^r)$ respectively $(\mathbf{I}_m, \mathbf{I}_m^r)$ is minimized. Within the method proposed by (Loop and Zhang, 1999) the projective distortion in the rectified images is minimized. Despite approaches of the methods are quite different, original $\mathbf{I}_b$ and $\mathbf{I}_m$ are warped such that epipolar lines are horizontal and an arbitrary object point $\mathbf{X}_i$ is mapped to the rectified image planes of $\mathbf{I}_b^r$ and $\mathbf{I}_m^r$ possessing the same y-coordinates, therefore

$$\mathbf{x}_b^r(x_b^r, y^r, 1) = \mathbf{x}_m^r(x_m^r, y^r, 1). \tag{3}$$

Recall that within the operation of rectification only the orientation of image planes are modified, the optical centers $\mathbf{C}$ and $\mathbf{C}^r$ remain identical. As a consequence also the optical rays and distances between object points and perspective centers are identical:

$$\begin{aligned} \mathbf{C}_m^r - \mathbf{X}_i &= \mathbf{C}_m - \mathbf{X}_i \\ \mathbf{C}_b^r - \mathbf{X}_i &= \mathbf{C}_b - \mathbf{X}_i. \end{aligned} \tag{4}$$

After deriving the homographies $\mathbf{H}_b$ and $\mathbf{H}_m$ gray values for pixels at integer positions in the rectified frames are calculated. Integer coordinates $\mathbf{x}_m$ are mapped to the original images using equation (1) and the respective gray values are interpolated.
The input of this module is the original imagery and the interior and exterior orientations. Within a first step radial distortion of the imagery is removed, then the actual rectification is carried out. Thereby an interface to common structure from motion and aerial triangulation software (as Bundler, VSFM, Inpho...) is provided. The output is epipolar images and the corresponding interior and exterior orientations.

2.2 Dense Stereo Matching Module

In this section the implemented module for dense image matching is described. It is a stereo method based on SGM but extends the classic approach as proposed in (Hirschmüller, 2008) by dynamically estimated disparity search ranges. Key advantages are reduced processing time, reduced memory consumption and the ability of processing scenes without previous knowledge about depth or disparity ranges. Furthermore, ambiguities of photo consistency measures as a result of weak or high frequent texture are resolved. However, a processing mode using the classical SGM approach operating on constant disparity search ranges is provided. Three different types of cost functions are implemented within the presented framework. In our experience the 9×7 Census cost (Zabih and Woodfill, 1994) is the most insensitive to parametrization, provides acceptable computation times and memory consumption and yields robust results. Moreover, local costs can be computed based on Mutual Information (P. Viola, 1997) and the DAISY descriptor (Tola et al., 2008).
All parameters and options can be specified by the user in an ASCII control file which is parsed at program start. The input of this module is rectified images as derived from the previous rectification module. The output is a raster data set representing the disparity of each base image pixel wit respect to the match image.

2.2.1 Review of the SGM algorithm

The problem of dense stereo matching is densely finding corresponding pixels across two views representing the same world object. Using epipolar images, potential correspondences (representing the same world object) are located in the same row of I_b and I_m and the problem can be reformulated as finding the disparity $d = x_m - x_b$. The SGM algorithm aims to estimate disparities across stereo pairs such that the global cost function

$$\begin{aligned} E(\mathbf{D}) = \sum_{\mathbf{x}_b} (C(\mathbf{x}_b, \mathbf{D}(\mathbf{x}_b)) \\ + \sum_{\mathbf{x}_N} P_1 T[\|\mathbf{D}(\mathbf{x}_b) - \mathbf{D}(\mathbf{x}_N)\| = 1] \\ + \sum_{\mathbf{x}_N} P_2 T[\|\mathbf{D}(\mathbf{x}_b) - \mathbf{D}(\mathbf{x}_N)\| > 1]. \end{aligned} \tag{5}$$

is minimized. Thereby $\mathbf{D}$ represents the disparity image holding disparity estimations of all base image pixels $\mathbf{x}_b$. T is an operator evaluating to one if the subsequent condition is true and evaluates to zero else. $\mathbf{x}_N$ denote base image pixels in the neighborhood of $\mathbf{x}_b$. The global cost function E is composed of a data term and two terms claiming for smooth surfaces. The data term is computed by pixel-wise similarity measures $C(\mathbf{x}_b, \mathbf{x}_m)$. The penalty parameters P_1 and P_2 control the gain of surface smoothing. Within a first step of the SGM method local costs $C(\mathbf{x}_b, d)$ for each base image pixel and its set of potential correspondences are calculated. Thereby d is an integer value in a constant range $d[d_{min}, d_{max}]$ defining all the potential correspondences. Each $C(\mathbf{x}_b, d)$ is assigned to a three dimensional, cube-shaped cost structure of the dimensions $r \times c \times (d_{max} - d_{min} + 1)$. Next, the aggregated costs $S(\mathbf{x}_b, d)$ are computed. Therefore C are

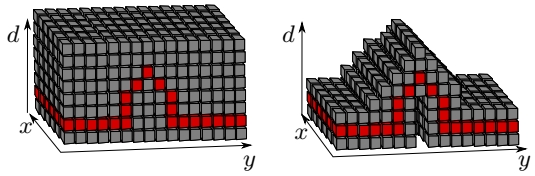


Figure 2: Cost structures of classic SGM (left) and tSGM (right). Red cubes represent costs for the true correspondences. Gray cubes mark the costs of potential correspondences, thus the disparity search ranges.

recursively accumulated along i image paths $\mathbf{r}_i$ according to

$$\begin{aligned} L_{\mathbf{r}_i}(\mathbf{x}_b, d) =& C(\mathbf{x}_b, d) + min(L_r(\mathbf{x}_b - \mathbf{r}_i, d), \\ & L_{\mathbf{r}_i}(\mathbf{x}_b - \mathbf{r}_i, d-1) + P_1 \\ & L_{\mathbf{r}_i}(\mathbf{x}_b - \mathbf{r}_i, d+1) + P_1, \\ & L_{\mathbf{r}_i}(\mathbf{x}_b - \mathbf{r}_i, i) + P_2) \\ & - min_k L_r(\mathbf{x}_b - \mathbf{r}_i, k) \end{aligned} \tag{6}$$

The last subtraction guarantees that $L_{\mathbf{r}_i}(\mathbf{x}_b, d) < C_max(\mathbf{x}_b, d) + P_2$. The sum over all paths

$$S(\mathbf{x}_b, d) = \sum_{\mathbf{r}_i} L_{\mathbf{r}_i}(\mathbf{x}_b, d) \tag{7}$$

results in a three dimensional structure holding costs for each pixel and its set of potential correspondences. Computing the minimum $d_{final} = min_d S(\mathbf{x}_b, d)$ for each $\mathbf{x}_b$ results in the final disparity image $\mathbf{D}$ minimizing an approximation of functional (5).

2.2.2 Modifications of the SGM algorithm - tSGM Within (Hirschmüller, 2008) a hierarchical approach was proposed to initialize and refine the MI matching cost. Initial disparity images were computed by matching high level (low resolution) image pyramids. The resulting disparities were then used to refine the MI matching cost for processing the subsequent pyramid level. Within our implementation disparities from a pyramid level l are furthermore used to limit the disparity search range used for matching the next lower pyramid level $l-1$. This hierarchical approach is carried out for all implemented matching costs. The search ranges are determined for each base image pixel individually. Let $\mathbf{D}^l$ be a disparity image resulting from matching the image pyramid l. For each pixel $\mathbf{x}_b$ the new search range is determined by evaluation of valid disparities around $\mathbf{D}^l(\mathbf{x}_b)$. If $\mathbf{x}_b$ was matched successfully minimum and maximum disparities d_{min} and d_{max} contained in a rather small 7×7 window are derived and stored in the two additional images $\mathbf{R}^l_{min}$ and $\mathbf{R}^l_{max}$. If $\mathbf{x}_b$ was not matched successfully a larger 31×31 window is searched for valid disparity estimations d_{min} and d_{max}. Moreover, the disparity estimation for $d(\mathbf{x}_b)$ of the current level l is updated to the median value of all disparities contained in the search window. The maximal disparity search ranges for valid and invalid pixels are limited to values of 16 and 32. In a next step the images $\mathbf{D}^l$, $\mathbf{R}^l_{max}$ and $\mathbf{R}^l_{min}$ are upscaled. These images define the disparity search range for matching images of the next pyramid level $l-1$. Potential correspondences during matching level $l-1$ are only searched in the ranges $[2*(x_b+d-d_{min}), 2*(x_b+d+d_{max})]$. Note that this implies a limitation of final search ranges to 32 pixels for valid and 64 pixels for invalid pixels. When processing the first (highest) pyramid level no initial disparity estimations are available. In this case all pixels in the match image along the horizontal epipolar are treated a potential correspondences. By the pixel-wise adaption of disparity search ranges the cubic shape of arrays holding the local costs $C(\mathbf{x}_b, d)$ and $S(\mathbf{x}_b, d)$ is no longer guaranteed (figure 2). In disparity space these structures represent a band containing potential disparities of the assumed surface. In practice all values of these structures are stored subsequently in one dimensional arrays and cost strings associated with a base image pixel are accessed using an image providing the respective offsets. Furthermore the path accumulation as given in equation (6) had to be modified. Since cost strings of neighboring pixels may overlap only partly or do not overlap at all, the terms $L_r(\mathbf{x}_b - \mathbf{r}_i, d+k)$ might not exist. In this case the bottom or top elements of the neighboring cost string $L_{\mathbf{r}_i}(\mathbf{x}_b - \mathbf{r}_i, d_{min}(\mathbf{x}_b - \mathbf{r}_i))$ and $L_{\mathbf{r}_i}(\mathbf{x}_b - \mathbf{r}_i, d_{max}(\mathbf{x}_b - \mathbf{r}_i))$ are employed. Equation (6) is enhanced according to

$$\begin{aligned} &\textbf{if} \quad d > d_{max}(\mathbf{x}_b - \mathbf{r}_i): \\ &\quad \bar{L}_{\mathbf{r}_i}(\mathbf{x}_b, d) = C_{\mathbf{r}_i}(\mathbf{x}_b, d) + P2 \\ &\textbf{if} \quad d < d_{min}(\mathbf{x}_b - \mathbf{r}_i): \\ &\quad \bar{L}_{\mathbf{r}_i}(\mathbf{x}_b, d) = C_{\mathbf{r}_i}(\mathbf{x}_b, d) + P2 \\ &\textbf{else}: \\ &\quad \bar{L}_{\mathbf{r}_i}(\mathbf{x}_b, d) = L_{\mathbf{r}_i}(\mathbf{x}_b - \mathbf{r}_i, d) \end{aligned} \tag{8}$$

Within matching of an image pair for a certain pyramid level the roles of base and match images are exchanged. This allows for a consistency check of estimated disparities, claiming $\|d_b - d_m\| \leq 1$. Moreover, speckles are filtered using an algorithm distributed by the OpenCV library (Bradski, 2000).
The penalty parameter P_2 adapts smoothing based on the gray values in the base image. In SURE P_2 possesses binary character and is calculated based on an canny edge image. Whereas if an edge was detected, low smoothing using $P_2 = P_{21}$ is applied. Increased smoothing is forced by setting $P_2 = P_{21} + P_{22}$ if no edge was detected. In addition to search range limitation, matching is carried out only on image areas representing scene parts commonly captured in the two views. In order to derive these image parts disparity maps of the lowest pyramid level are filtered rigorously. Afterwards pixels are passed column-wise from the left and right image borders to the image center. All pixels passed before the first successfully matched pixel is detected are invalidated and excluded from further processing. As the disparity range limitation this leads to a significant speed up of the matching procedure.

2.3 Structure Computation Module

In this paragraph the implemented algorithms for 3D object point triangulation are described. The input of the triangulation module are orientations of rectified/original base and match images and the correspondent disparity images. The output is a 3D point cloud or a depth image. Two main processing strategies are implemented. The first strategy directly computes the depths of single stereo pairs. For aerial applications where 2.5D DSMs are of interest this first approach is sufficient most of the times. To remove blunders and increase precision of point clouds derived from the single stereo models, all points are assigned to a ground aligned xy-grid and height values are median filtered. For close range and oblique aerial applications the second disparity map fusion approach is of larger importance. Dealing with real 3D structure, the griding approach and involved filter mechanisms are not applicable because too much information would be lost. In order to still remove blunders and improve accuracy when dealing with 3D scenes, redundant measurements across stereo models sharing a common base view are linked and checked for geometric consistency. Because for each set of redundant disparities estimations depth information has to be computed and time efficient algorithm has to be provided.

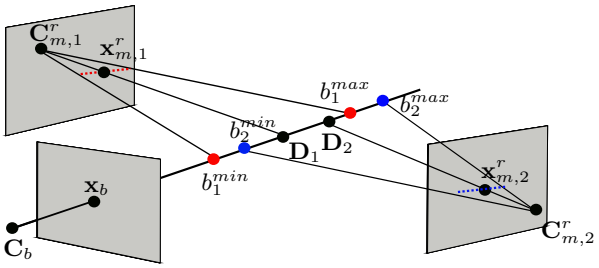


Figure 3: Confidence intervals of disparity estimations (blue and red dotted lines) induce a ranges on the base image ray $[b_n^{min}, b_n^{max}]$. If these ranges overlap disparity estimations are considered consistent.

2.3.1 Structure from Stereo Pairs 3D information for a pair of epipolar images $\mathbf{I}^r_b$, $\mathbf{I}^r_m$ (as generated during rectification described in 2.1) can be extracted using the well known formula (Kraus, 1994)

$$Z = \frac{Bf}{d}, \tag{9}$$

the so called normal-case of stereo imagery. Beside low computational costs precision analysis is convenient. Thereby B denotes the baseline $\|\mathbf{C}^r_b - \mathbf{C}^r_m\|$, d is the disparity and f represents the focal length. Z represents the z-component of the point with respect to the camera frames. 3D coordinate computation for the more general case in which varying f^r_x, f^r_y and sheering is present can easily be adapted by transforming the homogeneous image coordinates $\mathbf{x}_b$ to the plane $z = 1$, denoted by $(\bar{x}^r_b, \bar{y}^r_b, 1)$. Z-coordinates with respect to the first camera can then be calculated analogously to equation (9) setting $f = 1$. Respective x- and y coordinates can be obtained using the intercept theorem. The distance D^r_b between camera center $\mathbf{C}^r_b$ and object point $\mathbf{X}^r_b$ on the optical ray can be calculated as

$$D^r_b = \frac{B\sqrt{(\bar{x}^r_b)^2 + (\bar{y}^r_b)^2 + 1}}{d} \tag{10}$$

2.3.2 Structure from Multiple Stereo Pairs In many reconstruction scenarios captured imagery may overlap to a high degree. Therefore image pairs incorporated in the matching process can be chosen such that redundant disparity estimations for the same surface area are available. In a first step this redundancy is exploited to remove erroneous disparity estimations by evaluation of geometric consistency. Once a set of consistent measurements is derived final object point coordinates are computed. Thereby redundancy is further exploited to increase the precision of triangulated points. An example for two redundantly estimated disparities is visualized in figure (3). Correspondent image coordinates across n stereo models are derived as follows. Base image pixel coordinates $\mathbf{x}_b = (x_b, y_b, 1)$ are transformed to rectified base image coordinates $\mathbf{x}^r_b = (x^r_b, y^r_b, 1)$ using homographies as stated equation (2). For $\mathbf{x}^r_b$ disparities were calculated within the matching process and coordinates in the rectified match images can be obtained as $\mathbf{x}^r_m = (x^r_b + \mathbf{D}(\mathbf{x}^r_b), y^r_b, 1)$. Since $\mathbf{x}^r_b$ in general are real valued numbers the actual disparities are bilinearly interpolated. The distance D^r_b between the optical centers of rectified base images and the object point can be efficiently calculated using equation (10). Linking a base image pixel $\mathbf{x}_b$ with n stereo models results in n depth estimations $(D^r_{b,1}, ..., D^r_{b,N})$. Note that $D^r_{b,n}$ are defined with respect to the rectified base coordinate systems. However, equation (4) clarifies that depths $D^r_{b,n}$ along rectified base image rays equal the depths along the original base image rays therefore $D^r_{b,n} = D_{b,n}$. By extension of equation (10) depths in the common original base coordinate system can be calculated as

$$D_{b,n}(\mathbf{x}_b, \mathbf{T}_n, \mathbf{D}_n) = \frac{B\sqrt{(\mathbf{t}_{1,n}\mathbf{x}_b)^2 + (\mathbf{t}_{2,n}\mathbf{x}_b)^2 + 1}}{\mathbf{D}_n(\mathbf{T}_n\mathbf{x}_b)}. \tag{11}$$

Thereby $\mathbf{T}_n = \mathbf{K}^{r\,-1}_b\mathbf{H}_b$ and $\mathbf{t}_{1,n}$, $\mathbf{t}_{2,n}$ denote the first respectively second row of $\mathbf{T}_n$.

Outlier elimination In the dense stereo matching process erroneous disparity estimations are eliminated using forward-backward consistency check and speckle filters. However, not all mismatches can be removed by these 2D filter methods. Therefore erroneous disparities are filtered additionally by checking for geometric consistency in object space. It is claimed that 3D coordinates implied by redundant disparities across a set of stereo models are spatially consistent within some confidence interval. For the special case of rectified images the consistency check of 3D points can be reduced to a one dimensional problem. This allows for fast processing and exact error modeling. Let a base image be rectified and matched against n match images. Furthermore it is assumed that disparities are estimated with an precision defined by a confidence interval σ along the epipolar line. The interval σ induces an uncertainty range $R_n = [b_n^{min}, b_n^{max}]$ on the optical base image ray defined by $\mathbf{x}_b$. Its borders are calculated according to equation (11) as

$$b_n^{min,max} = D_{b,n}(\mathbf{x}_b, \mathbf{T}_n, \mathbf{D}_n(\mathbf{x}_b) \pm 0.5\sigma) \tag{12}$$

If the uncertainty ranges R_n of the single object points are overlapping, the depth measurements are regarded as consistent and assigned to a cluster. All measurements contained by the biggest cluster are then used for the final object point triangulation. If two or more clusters possess the same size m, the cluster providing the lowest average of ray intersection angles

$$\frac{1}{m}\sum_m (\angle(\mathbf{X} - \mathbf{C}_b, \mathbf{X} - \mathbf{C}_m)) \tag{13}$$

is considered as most reliable and used for structure computation. Note that within this approach image space accuracies are correctly propagated. This is important for reliable outlier detection particularly in presence of varying geometric configurations of stereo models.

Triangulation The problem of 3D point triangulation minimizing the reprojection error is a nonlinear problem. Typically it is solved using iterative numerical approaches as Gauss-Newton or Levenberg-Marquardt. This involves solving a linear system of equations possessing a design matrix $\mathbf{A}$ with two rows per incorporated model. In the special case of rectified images the problem can be reformulated as a system of equations possessing an $\mathbf{A}$ with only one row per stereo model. In projective space object point triangulation can be formulated as linear problem (R. I. Hartley, 2004). However, only an algebraic error without any geometric meaning is minimized. Within SURE a method for 3D coordinate computation by minimization of the object space error from multiple redundant depth

$$\sum_n (\hat{D} - D_m)^2 \stackrel{!}{=} min \tag{14}$$

is implemented. The solution is simply the average of estimated depths $\frac{1}{n}\sum D_n$. The accuracy along the optical ray can be estimated using standard deviations. Despite this method is fast, geometric properties of different image pairs are not properly modeled. Therefore an approach minimizing the reprojection error in

the rectified match images

$$\sum_n \frac{1}{2}(\hat{x}_m - x_m)^2 \stackrel{!}{=} min \tag{15}$$

is provided. Reprojection errors can be expressed by scalars since measurements as well as updated image coordinates are located on the horizontal epipolar line. The minimum of equation (15) is defined by derivation and equating to zero. Using equation (10) and the relation $D_m = \hat{D}$ this functional can be reformulated as function dependent of the common unknown depth $\hat{D}$

$$f_m(\hat{D}) = \frac{B_m\sqrt{(\bar{x}^r_{b,m})^2 + (\bar{y}^r_{b,m})^2 + 1}}{\hat{D}} - d_m. \tag{16}$$

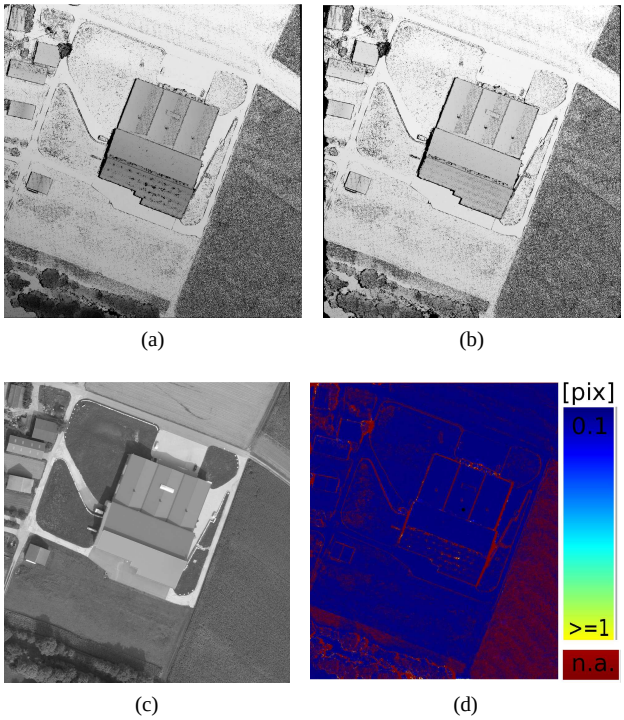
For n stereo models this leads to a set of n equations nonlinear in $\hat{D}$. The optimal $\hat{D}$ minimizing equation (15) is determined using the Levenberg-Marquardt (Lourakis, Jul. 2004) or Gauss-Newton algorithm. This implies linearization and solving a inhomogeneous linear system $\mathbf{Ax} = \mathbf{b}$. The least squares solution is obtained as $\hat{\mathbf{x}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}$. Note that since the design matrix $\mathbf{A}$ is of dimension $n \times 1$, $\mathbf{A}^T\mathbf{A}$ is a scalar, and no matrix inversion is required. Initial depth values are calculated from equation (14). Accuracies of estimated depths can be obtained by evaluation of covariance matrices. So far we assumed accuracies of disparities estimations to be identical. Using a-priori knowledge of matching accuracies in dependence of ray intersection angles, outlier detection could be refined and weighted adjustment could be used within the minimization of the reprojection error.

3 RESULTS

3.1 Comparison of SGM and tSGM

Within this section results of SGM and tSGM approaches are compared. Therefore two sets of rectified images pairs were matched and resulting disparity images were evaluated. For the SGM solution a constant disparity search range covering exactly all prevalent disparities was specified. The first image pair consists of two (2298×2290) sub tiles cropped from two large format aerial frames. Matching was carried out on full resolution imagery. The resulting disparity image was calculated in 44 seconds(Pentium R dual core, 2.6 GHz) and is shown in figure 4a. The maximal memory consumption amounted 2.6GB. The parallax image derived by tSGM was computed in 30 seconds and is displayed in figure 4b. A visualization of dynamic search ranges for all subsequent pyramid levels is displayed in figure 5. Due to the reduced size of the structures used for cost computation and cost aggregation memory consumption of tSGM could be reduced by 68.2% to 0.8GB. Same observation holds for processing times. Within this example execution time was reduced by 31.8%. Note that for the chosen aerial scenario minimal and maximal prevalent disparities do not heavily vary and cube structures are comparable small. For scenes inducing larger variances in depth, as particularly prevalent in close range applications, memory demand can be reduced by multiples. For a second test two images from the Fountain data set (Strecha et al., 2008) were rectified and matched. Matching using the classical approach was carried out in 65 seconds (I7 quad core, 3.4 GHz). The top memory consumption amounted 21.1 GB. The time for matching using the tSGM solution could be reduced by 89.3% to 6.88 seconds. The memory consumption could be reduced by 93.8% to 1.3 GB which enables processing on standard computers. However, in all tests SGM was calculated using the same core algorithmic for cost computation and aggregation as used for tSGM. Assuming regular cubic structures, before mentioned operations could be designed and executed more efficiently for the classical SGM and lead to lower processing times. The memory consumption of tSGM can be further reduced by tiled processing. Thereby tile sizes are adapted according to the available physical memory.

Visual comparison of the disparity images clarifies that tSGM hinders reconstruction of largely undulating structures represented by only few pixels in the images. As for the power pole in figure 4d small pixel patches might not be passed to lower pyramid levels due to resolution reduction and smoothing. Therefore the predicted search range in the next higher resolution pyramid might not contain the correct disparities and reconstruction for these objects might fail. However, in many data sets the surfaces are captured in various angles. Structures therefore might be represented by a larger number of pixels in additional views, which then enables successful reconstruction. Moreover, the proposed tSGM algorithm provides beneficial reconstruction of low textured objects and objects possessing repetitive texture as the roof in figure 4c. High frequencies are not passed to lower levels which enables robust parallax estimation. In subsequent levels ambiguities are resolved due to the reduced search range which leads to a reduction of mismatches. The same observation holds for image parts possessing weak texture and larger differences in appearance as the ground in the Fountain data set. When matching low pyramid levels the appearance is more similar and Census matching costs are more distinctive since a larger area in object space is captured by the 9×7 correlation window. As before, disparities are propagated to subsequent levels and ambiguities are resolved by the limited disparity search range. This leads to a higher completeness of the disparity maps.



(a) (b)

(c) (d)

Figure 4: (a): Disparity map of classic SGM approach. (b): Disparity map of tSGM, (c): Original image (d): Absolute differences in disparity maps (a) and (b)

Figures 4d and 6d decode the absolute differences of the SGM and tSGM disparity images. These differences could only be cal-

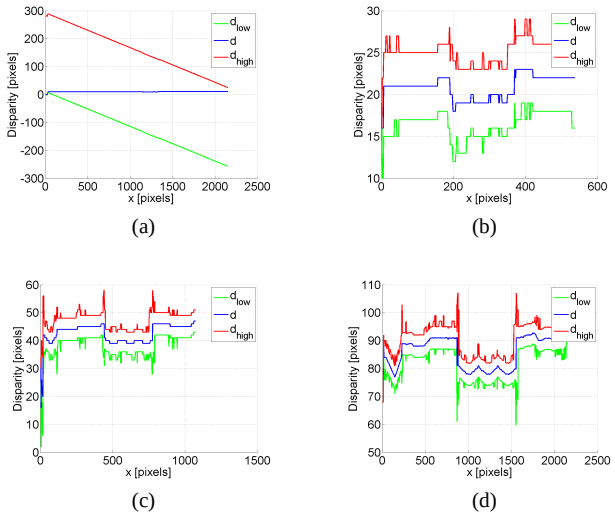


Figure 5: Visualization of disparity search ranges of pyramid levels 3-0 (a-d). Blue line marks the estimated / interpolated disparity from previous pyramid level. Green and red mark the disparity search range for current level.

culated for pixels for which disparity estimations in both maps were available. If a disparity was estimated only in one or none of the maps the pixel is marked in red. Disparities differing more than one pixel are mainly located at larger disparity steps. Visual inspection leads to the conclusion that within the SGM approach edges are reconstructed more clearly, particularly for the close range example. These errors propagate and lead to differences on sub-pixel level in surrounding edge areas. However, on continuous surfaces both solutions yield rather same results and the differences of disparities are mostly below 0.1 pixels.

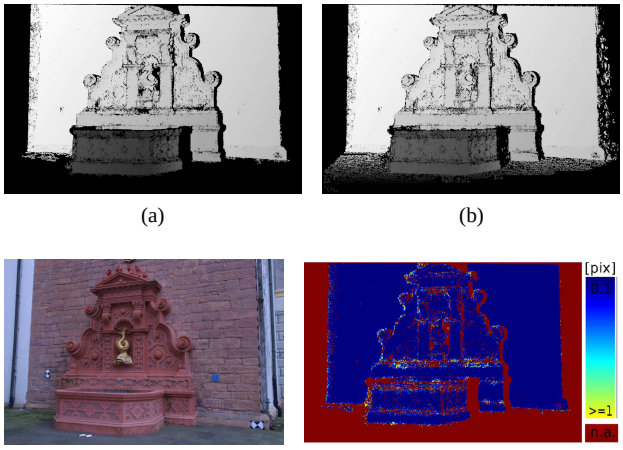


Figure 6: (a): Disparity map of classic SGM approach. (b): Disparity map of tSGM, (c): Original image (d): Absolute differences in disparity maps (a) and (b)

3.2 Robustness Regarding Parameters for Variety of Scenes

Insensitivity to parametrization is essential to reduce extensive interference by the operator. In order to show that the algorithm is rather robust to parametrization, datasets captured by different type of cameras and varying geometric configurations were processed using the same matching settings. Within the 3D point triangulation the only parameter changed was the minimal number of geometric consistent depth estimations required for an object point to be evaluated as valid. Figure 7 shows the resulting point clouds/meshes. The first dataset in 7a was captured with a large format aerial camera (UltraCamX). Side and forward overlap amount 80/70% possessing a GSD of 8cm. The second dataset 7b was captured by a unmanned areal vehicle using a consumer grade camera (Haala and Rothermel, 2012b). The overlap amounts 75/70% at a GSD of 4 to 8cm. The well documented Fountain dataset (Strecha et al., 2008) provided by the EPFL is shown in figure 7d. Within processing each image was matched against at least 4 proximate images leading to 27 stereo models and 54 disparity maps. The time for dense matching amounted less then 3 minutes including input and output operations (i7 quad core, 3.4 GHz). Multi-view triangulation was carried out in 63 seconds including IO operations. The dataset displayed in figure7c was collected using the camera of a mobile phone (HTC One S). The captured object is a sculpture (approx. 1.5m in height) captured in unbeneficial light conditions. Despite the signal-to-noise ratio of the mobile phone imagery is rather low, dense surfaces could be reconstructed in most parts. In figure 7e a reconstruction of the test object 'Testy' (35cm in height) is visualized. Figure 7f shows the reconstruction of oblique UAV imagery. Thereby 198 frames were extracted from a video sequence kindly provided by Fraunhofer IOSB. All displayed point clouds are direct output of the algorithm and no further point cloud processing was applied.

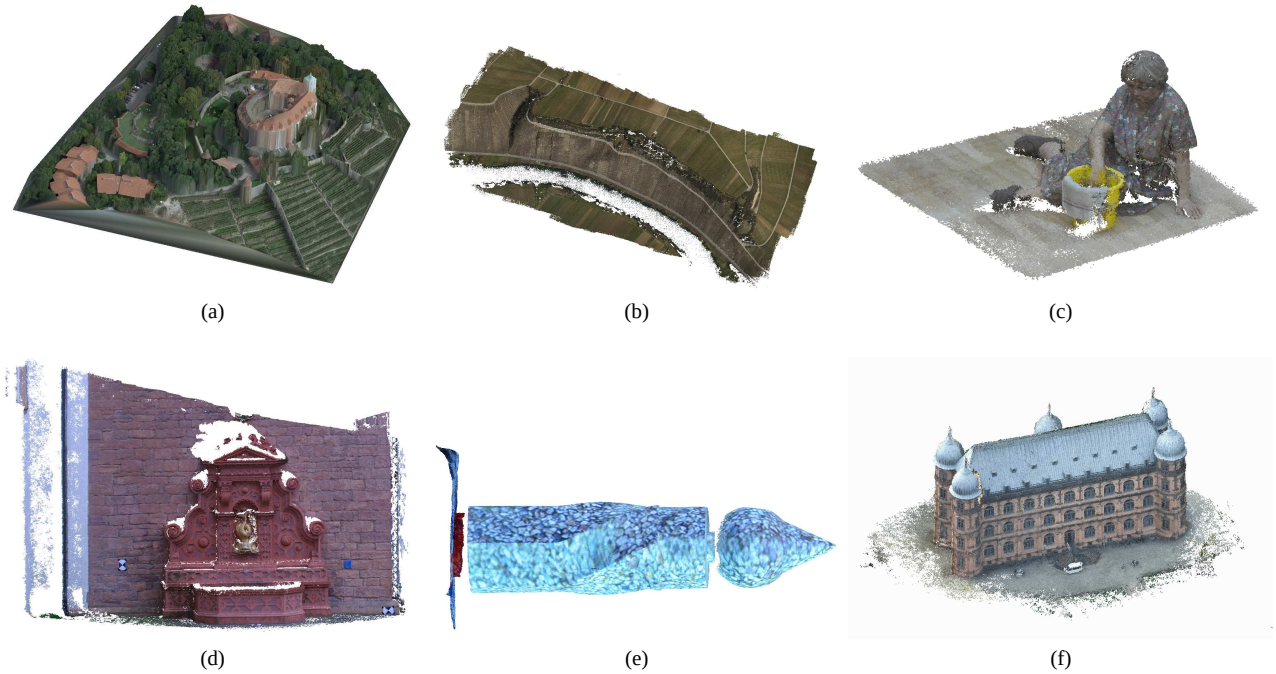
4 SUMMARY

Within this article implementation details of the software package SURE were presented. One main contribution is the enhancement of the SGM approach by the capability of searching pixel correspondences using dynamic disparity search ranges. The formerly cube shaped structure storing costs of potential correspondences in a constant search range was modified to a tube-shape structure containing costs from dynamic search ranges. Thereby the path aggregation was enhanced such that same global cost function as given in the classic SGM approach is minimized. The second contribution is the exploitation of epipolar geometry for multi view structure computation and blunder filtering. The problem was formulated as a nonlinear problem minimizing the reprojection error in dependence of only in the depth. This set of equations can be solved using iterative numerics, for which no matrix inversion is needed. At the same time blunder filtering based on the geometric consistency of disparity estimations can be reduced to a one dimensional clustering problem. This strategy of fusing disparity maps leads to depth maps with reduced number of outliers and increased precision. Compared to the classic SGM approach within our tests memory demands as well as computation times could be reduced by close to 90%. Moreover, the completeness of results was increased. Height discontinuities were not as clearly reconstructed as in the classical approach. The algorithm scales well to large number of images and high resolution imagery. This and robustness regarding parametrization makes it suitable for reconstruction of close range, UAV and aerial imagery. The SURE package as well as the libTSgm library providing an OpenCv API is available for free and non-commercial use.

REFERENCES

A. Gruen, E. B., 1988. Geometrically constrained multiphoto matching. Photogrammetric Engineering and Remote Sensing 54, pp. 633–641.

Bradski, G., 2000. The opencv library. Dr. Dobb's Journal of Software Tools.



(a) (b) (c)

(d) (e) (f)

Figure 7: Exemplary point clouds derived by SURE using identical parameters for dense matching. For object point triangulation one parameter for the minimal number of consistent was adapted for the respective image configuration. No additional filtering was applied to point clouds and meshes. Details on the datasets are given in section 3.2.

Curless, B. and Levoy, M., 1996. A volumetric method for building complex models from range images. In: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques, ACM, pp. 303–312.

Fuhrmann, S. and Goesele, M., 2011. Fusion of depth maps with multiple scales. In: Proceedings of the 2011 SIGGRAPH Asia Conference, SA '11, ACM, New York, NY, USA, pp. 148:1–148:8.

Furukawa, Y. and Ponce, J., 2010. Accurate, dense, and robust multiview stereopsis. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(8), pp. 1362–1376.

Furukawa, Y., Curless, B., Seitz, S. and Szeliski, R., 2010. Towards internet-scale multi-view stereo. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pp. 1434–1441.

Fusiello, A., Trucco, E. and Verri, A., 2000. A compact algorithm for rectification of stereo pairs.

Gehrig, S., Eberli, F. and Meyer, T., 2009. A real-time low-power stereo vision engine using semi-global matching. In: M. Fritz, B. Schiele and J. Piater (eds), Computer Vision Systems, Lecture Notes in Computer Science, Vol. 5815, Springer Berlin Heidelberg, pp. 134–143.

Goesele, M., 2007. Multi-view stereo for community photo collections.

Goesele, M., Curless, B. and Seitz, S., 2006. Multi-view stereo revisited. In: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, Vol. 2, pp. 2402–2409.

Haala, N. and Rothermel, M., 2012a. Dense multi-stereo matching for high quality digital elevation models. PFG Photogrammetrie, Fernerkundung, Geoinformation 2012(4), pp. 331–343.

Haala, N. and Rothermel, M., 2012b. Dense multiple stereo matching of highly overlapping uav imagery. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XXXIX-B1, pp. 387–392.

Hermann, S. and Klette, R., 2012. Evaluation of a new coarse-to-fine strategy for fast semi-global stereo matching. In: Proceedings of the 5th Pacific Rim conference on Advances in Image and Video Technology - Volume Part I, PSIVT'11, Springer-Verlag, Berlin, Heidelberg, pp. 395–406.

Hermann, S. and Klette, R., 2013. Iterative semi-global matching for robust driver assistance systems. In: K. Lee, Y. Matsushita, J. Rehg and Z. Hu (eds), Computer Vision ACCV 2012, Lecture Notes in Computer Science, Vol. 7726, Springer Berlin Heidelberg, pp. 465–478.

Hirschmüller, H., 2008. Stereo processing by semiglobal matching and mutual information. IEEE Transactions on Pattern Analysis and Machine Intelligence 30, pp. 328–341.

Hirschmüller, H., Ernst, I. and Buder, M., 2012. Memory efficient semi-global matching. International Annals of Photogrammetry and Remote Sensing.

Koch, R., Pollefeys, M. and Gool, L. J. V., 1998. Multi viewpoint stereo from uncalibrated video sequences. In: Proceedings of the 5th European Conference on Computer Vision-Volume I - Volume I, ECCV '98, Springer-Verlag, London, UK, UK, pp. 55–71.

Kraus, K., 1994. Photogrammetrie - Band 1. Ferd. Dümmlers Verlag, ISBN: 3-427-78645-5.

Loop, C. and Zhang, Z., 1999. Computing rectifying homographies for stereo vision. In: Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on., Vol. 1, pp. 2 vol. (xxiii+637+663).

Lorensen, W. E. and Cline, H. E., 1987. Marching cubes: A high resolution 3d surface construction algorithm. In: Proceedings of the 14th annual conference on Computer graphics and interactive techniques, SIGGRAPH '87, ACM, New York, NY, USA, pp. 163–169.

Lourakis, M., Jul. 2004. levmar: Levenberg-marquardt nonlinear least squares algorithms in C/C++. [web page]

http://www.ics.forth.gr/~lourakis/levmar/. [Accessed on 31 Jan. 2005.].

Merrell, P., Akbarzadeh, A., Wang, L., michael Frahm, J. and Nistr, R. Y. D., 2007. Real-time visibility-based fusion of depth maps. In: In Int. Conf. on Computer Vision and Pattern Recognition.

Okutomi, M. and Kanade, T., 1993. A multiple-baseline stereo. IEEE Trans. Pattern Anal. Mach. Intell. 15(4), pp. 353–363.

P. Viola, Wells III, M. W., 1997. Alignment by maximization of mutual information. International Journal of Computer Vision 24, pp. 137–154. 10.1023/A:1007958904918.

Pierrot-deseilligny, M. and Paparoditis, N., 2006. A multiresolution and optimization-based image matching approach: An application to surface reconstruction from spot5-hrs stereo imagery. In: In: Proc. of the ISPRS Conference Topographic Mapping From Space (With Special Emphasis on Small Satellites), ISPRS.

R. I. Hartley, A. Z., 2004. Multiple View Geometry in Computer Vision. Second edn, Cambridge University Press, ISBN: 0521540518.

Remondino, F. and Zhang, L., 2006. Surface reconstruction algorithms for detailed close range object modelling. In: International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. XXXVI, part 3, pp. 117–123.

Roy, S. and Cox, I. J., 1998. A maximum-flow formulation of the n-camera stereo correspondence problem. In: Proceedings of the Sixth International Conference on Computer Vision, ICCV '98, IEEE Computer Society, Washington, DC, USA, pp. 492–.

Seitz, S. M., Curless, B., Diebel, J., Scharstein, D. and Szeliski, R., 2006. A comparison and evaluation of multi-view stereo reconstruction algorithms.

Strecha, C., von Hansen, W., Van Gool, L., Fua, P. and Thoennessen, U., 2008. On Benchmarking Camera Calibration and Multi-View Stereo for High Resolution Imagery.

Tola, E., Lepetit, V. and Fua, P., 2008. A fast local descriptor for dense matching. In: Conference on Computer Vision and Pattern Recognition, Alaska, USA.

Turk, G. and Levoy, M., 1994. Zippered polygon meshes from range images. In: Proceedings of the 21st annual conference on Computer graphics and interactive techniques, SIGGRAPH '94, ACM, New York, NY, USA, pp. 311–318.

Wenzel, K., Abdel-Wahab, M., Cefalu, A. and Fritsch, D., 2011. A multi-camera system for efficient point cloud recording in close range applications. In: LC3D workshop, pp. 37–46.

Zabih, R. and Woodfill, J., 1994. Non-parametric local transforms for computing visual correspondence. In: J.-O. Eklundh (ed.), Computer Vision ECCV '94, Lecture Notes in Computer Science, Vol. 801, Springer Berlin Heidelberg, pp. 151–158.