EFFICIENT, PRECISE, AND ACCURATE UTILIZATION OF THE UNIQUENESS CONSTRAINT IN SELF-CALIBRATED STEREO

Xenophon Zabulis[1], *Ugur Topay*[2] *and A. Aydin Alatan*[2]

[1]Informatics and Telematics Institute, Centre for Research and Technology Hellas
1st Km Thermi-Panorama Road, Thessaloniki, Greece, GR-57001
xenophon@iti.gr

[2]Dept. of Electrical-Electronics Eng., METU
Balgat 06531 Ankara TURKEY
ugur.topay@tubitak.gov.tr, alatan@eee.metu.edu.tr

ABSTRACT

In this paper, the depth cue due to the assumption of texture uniqueness is reviewed. The spatial direction over which a similarity measure is optimized, in order to establish a stereo correspondence, is considered and methods to increase the precision and accuracy of stereo reconstructions are presented. It is further presented that the proposed method is quite robust to projective distortions due to less accurate camera parameters, possibly obtained through self-calibration. An efficient implementation of the above methods is also offered, based on a scale-space treatment of the data. The above contributions are integrated in a generic and parallelizable implementation of the uniqueness constraint to observe speedup and increase in the fidelity of surface reconstruction.

1. INTRODUCTION

Despite the recent growth of the use of spectral information (color) in the stereo problem, the cue due to the uniqueness constraint remains relevant, as utilized by contemporary stereo systems [1]. Its locality of access to the data (images) facilitates multi-view and parallel implementations of the cue, for real-time applications such tele-immersion, 3D television, and telemedicine. Its advantage over other cues is its independence from silhouette extraction, assumption of cameras around the scene [2], or the same baseline [3], as well as spectral calibration (color photoconsistency approaches e.g. [4]). This work focused at the optimization of the use of the uniqueness constraint in the Lambertian and textured domains, and acknowledges its integration with photometric cues as a future goal.

Authors are grateful for support through the 3DTV European Network of Excellence, 6th Framework IST Programme.

The uniqueness constraint assumes that, at least locally, each locus on a surface is uniquely textured. Its point correspondences can be, thus, located by searching for its most similar depictions in the acquired images. A number of works compensate for the projective distortion in the matching process (or otherwise, match the textures in 3D), to obtain more matches and accuracy [5, 6, 7, 8, 9]. Treating the imaged surfaces as locally planar allows the backprojections of images at hypothetical planar patches and, in turn, the prediction that backprojections should match if the patch coincides with the surface. A match then provides the estimations of locus and orientation of the surface. In this work, it is shown that even the above compensation is subject to inaccuracy due to image discretization and methods to efficiently rectify are proposed.

The remainder of this paper is organized as follows. In Sec. 2 related work is reviewed. In Sec. 3 the method by which the cameras are self-calibrated is presented. The proposed methods on accuracy and precision are presented in Sec. 4. These methods are then computationally optimized in Sec. 5. In Sec. 6, contributions are compiled into a parallelizable multiview stereo algorithm, which is demonstrated, and discussed.

2. RELATED WORK

Correspondence estiblishment based on the uniqueness constraint is typically performed by similarity matching of (back) projections, through the optimization of a similarity criterion along a spatial direction. The points where the similarity metric is locally optimized are regarded as occupied. This work is focused at local methods as global optimization (e.g. [7, 10, 3, 11]) can yield local minima of the overall cost function and are not nessesarily paralelizable.

Either a single (as in epipolar stereo or plane sweeping

[12]) or multiple orientations of the surface may be considered in the optimization, thus classifying methods into two categories, referred to as *sweeping* and *orientation-search* respectively. Sweeping exhibits low complexity because a single orientation $\vec{\kappa}$ is evaluated, but is inaccurate at intense slants, wide baselines, and verging cameras. Orientation search is computationally more complex ($\vec{\kappa}$ is optimized), but provides with an estimation of the surface normal.

The assumption of surface continuity can be enforced to improve the quality of reconstruction. In epipolar stereo, some approaches to enforcing this constraint are to filter the disparity map [13], bias disparity values that are in coherence with neighboring [3], or require "inter-scanline" consistency [14]. An abundance of approaches for 3D filtering of the results exists in the "deformable models" literature (see [15] for a review). In this work, interest is focused in RBF-based methods [16] due to their ability to facilitate smooth interpolation in 3D. The approach proposed in Sec. 4.3 differs from [7, 10], in the explicit targeting of similarity's local maxima instead of implicitly by optimization of a global functional.

3. SELF-CALIBRATION OF THE CAMERA

Camera self-calibration is a quite powerful method considering the fact that one can obtain the camera parameters without using the 3-D structure of the scene. The camera self-calibration methods can be classified into two groups. In the first group, while the camera parameters for the Euclidean reconstruction are estimated directly by solving Kruppa Equations [17] or using the properties of the E-Matrix [18, 19] the camera parameters can also be estimated by starting from a projective reconstruction, as an alternative approach [20].

In this study, a method, which uses the properties of the E-Matrix, (rank-2 and equal singular values) is utilized. The relation between the fundamental matrix F and the camera calibration matrices (A_i, A_j where i and j are frame numbers) is given by $E_{ij} = A_j^T F A_i$. As a first step, the initial focal length α is estimated by assuming unity aspect ratio ($\alpha_u = \alpha_v$), as well as the principal point at the center of the image. Next, since the E-Matrix should have equal singular values, it is possible to decompose this matrix as follows [21]:

$$[1/b_2 \cdot b_3 \;\; 1/b_1 \cdot b_3 \;\; 1/b_1 \cdot b_2] \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = 0, \qquad (1)$$

where $E = [b_1, b_2, b_3]^T$. By using this property, a linear equation in terms of the unknown focal lengths for each image, can be obtained and then solved by using F-matrices F_{12} and F_{13}.

Compared to other focal length estimation methods [22, 23], this method has the advantage of utilizing all F-Matrices belonging to to ith image.

After initial estimation of the focal lengths for each image, with the assumption of utilizing the same camera for all views, the camera parameters can be estimated by the minimization of the following cost function:

$$\min \sum_{i=1}^{N} \kappa_i \left(1 - \frac{\sigma_2^i}{\sigma_1^i}\right) \qquad (2)$$

where N is equal to the number of F-Matrices, whereas $E^i = A^T F^i A = U diag(\sigma_1^i, \sigma_2^i, 1) V^T$. The cost function simply tries to equate to singular values with each other to obtain the other unknown camera parameters.

4. INCREASING ACCURACY & PRECISION

A generic formulation of applying the uniqueness cue in stereo is presented, in order to define methods for increasing the accuracy of the results. Both methods are based on the assumption of surface continuity.

4.1. Cue formulation

Let a calibrated image pair $I_{i=1,2}$, and a planar surface patch $\mathcal{S}$, of size is $\alpha \times \alpha$, centered at $\vec{p}$, with unit normal $\vec{n}$. Back-projecting I_i onto $\mathcal{S}$ yields images $w_i(\vec{p}, \vec{n})$:

$$w_i(\vec{p}, \vec{n}) = I_i \left(P_i \cdot \left(\vec{p} + R(\vec{n}) \cdot [x' \; y' \; 0]^T\right)\right), \qquad (3)$$

where P_i is the projection matrix of I_i, $R(\vec{n})[0\,0\,1]^T = \vec{n}$, and x', y' local coordinates on $\mathcal{S}$.

When $\mathcal{S}$ is tangent at a world surface, $w_i(\vec{p}, \vec{n})$ are identities of the surface pattern. Thus $I_1(P_1\vec{x}) = I_2(P_2\vec{x})$; $\forall \vec{x} \in \mathcal{S}$, and therefore their similarity is optimal. Otherwise, w_i are dissimilar because they are collineations from different surface regions. Thus surface loci and corresponding normals can be recovered by detecting the similarity local maxima (SLM) across space. To do so, function $\vec{V}(\vec{p}) = s(\vec{p})\vec{\kappa}(\vec{p})$, is evaluated as:

$$s(\vec{p}) = \max_{\vec{n}} \left(sim(w_1(\vec{p}, \vec{n}), w_2(\vec{p}, \vec{n}))\right), \qquad (4)$$

$$\vec{\kappa}(\vec{p}) = \arg\max_{\vec{n}} \left(sim(w_1(\vec{p}, \vec{n}), w_2(\vec{p}, \vec{n}))\right). \qquad (5)$$

where $s(\vec{p})$ the optimal correlation value at $\vec{p}$, and $\vec{\kappa}(\vec{p})$ the optimizing orientation, Metric sim can be SAD, SSD, NCC, MNCC etc (henceforth MNCC). To evaluate sim, a $r \times r$ sampling lattice of points is assumed on $\mathcal{S}$.

4.2. Accuracy

The detection of maxima by sweeping, exhibits decreased performance in the presence of slant. In Fig. 1(b), suppression of valid maxima occurs when $\vec{\kappa}$ is in wide disagreement with the surface normal, because $\vec{\kappa}$ may point to valid SLMs. When $\vec{n}$ is optimized (5) suppression attenuates, because $\vec{\kappa}$ does not point to neighboring surface elements, Fig. 1(c). As shown, despite matching of texture in 3D, the estimation of $\vec{\kappa}$ can be still significantly inaccurate.

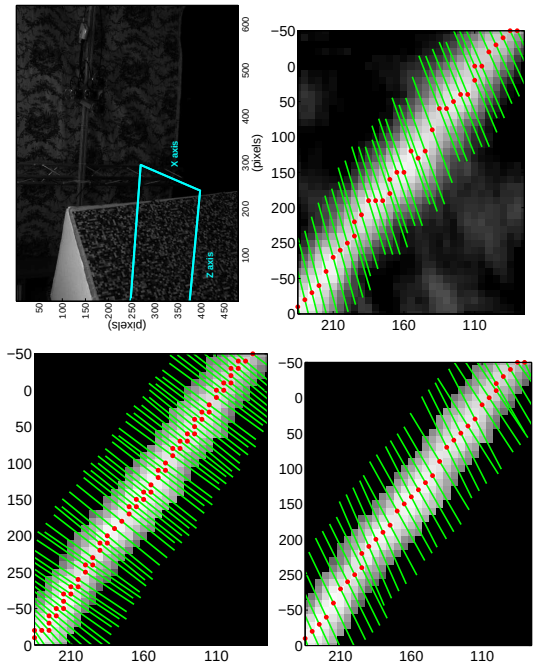

Fig. 1. SLM detection. Clockwise from top left: (a) Image from a binocular pair (baseline $156mm$), showing a horizontal slice at which $\vec{V}$ was calculated. Detection of SLMs with by sweeping (b) may give rise to spurious suppressions of valid SLMs e.g. at (180,200). Optimizing $\vec{\kappa}$ (c), attenuates the effect. Best results are obtained by linear filtering and recalculation of SLMs (d). In the graphs, red dots are SLMs, green lines are $\vec{\kappa}$, axes are in mms, horizontal is xx', vertical zz', voxel=$125mm^3$, $r = 21 \times 21$, $\alpha = 20mm$. In (a), checker size was $1cm^2$ and target was $\approx 1.5m$ from the cameras.

If surface continuity is enforced, SLMs are detected along a more accurate estimation of the normal. To do so, $\vec{V}$ was first estimated through (4), (5) and SLMs were detected along $\vec{\kappa}$ normals. Then $\vec{\kappa}$ was filtered and SLMs re-deteted along the new $\vec{\kappa}$. Directions in $\vec{\kappa}$ at occupied voxels, were replaced by the normal of the least-squares fitting plane through the neighboring occupied voxels. For empty voxels $\vec{p}_e$, $\vec{\kappa}(\vec{p}_e) = \sum_j \beta_i \vec{\kappa}(\vec{p}_j) / \sum \beta_j$, where j enumerated the occupied voxels within $\vec{p}_e$'s neighborhood. In the result Fig. 1(d), $\vec{\kappa}$ is more accurately estimated and a denser 3D reconstruction is provided. Note that $\vec{\kappa}$'s optimization may not be completely avoided, since the initial refinement of SLMs is required for robust application of the above method.

The observed inaccuracy of (5) is attributed to image discretization. $\mathcal{S}$'s projections cover different amounts of pixels n_i, which are analogous to the reciprocals of distance squared and relative slant of $\mathcal{S}$ to the cameras. When slant and/or distance increase, fewer image pixels are sampled, to evaluate (5). Then, the differences of fewer intersities typically exhibit less variance (or more "similarity") and, thus, a bias is observed towards greater slants and distances.

4.3. Precision

To increase precision to subvoxel s and ∇s are interpolated to detect the loci where $\nabla s = 0$ since they coincide with SLMs. A continuous interpolating surface is obtained, which is then discretized by a triangle mesh:

First, the initial (see Sec. 4.2) 1D search of the loci of SLMs along $\vec{\kappa}$ is refined by interpolating s along $\kappa(\vec{p})$. The polynomial $y = \sum_\kappa c_\kappa x^\kappa$ is solved as to c_κ, for fitting values $y_k = s(\vec{p} + x\vec{\kappa}(v))$, $x_k = -1, 0, 1$, and $k = 1, 2, 3$. The updated SLM locus is $\vec{p} - (\beta/2\alpha)\kappa(\vec{v})$. Second, the loci of SLMs are refined by virtue of the assumed continuity of $\vec{V}$. The pursued topological space S_f is given by the zero set of $S_f = f \in R^3 | f(p) = 0$.

Function f is instantiated as in [16], via a Radial Basis Function (RBF) framework and the use of pivot points (f's samples). Only off-surface pivot points, one "inside" and one "outside", the surface for each SLM $\vec{p}_m$ are selected. These points reside at $\vec{p}_m \pm \lambda\vec{\kappa}$. with values $\pm(\nabla s)(\vec{p}_n)$, respectively; λ is chosen less than voxel size (i.e. .8) to avoid interference with neighboring SLMs.

To extract S_f, f is calculated at a given resolution and inputted to the Marching Cubes algorithm [24] which outputs S_f as a triangular mesh. Fast evaluation of f is performed by taking only close surface pivots (e.g. 5-7 voxels distance) into account. Fig. 2 demonstrates the process and results.

5. SPEEDING-UP THE COMPUTATION

Two hierarchical, coarse-to-fine iterative methods are proposed for the accelaration of the search for SLMs. Sec. 5.1 concerns the estimation of $\vec{\kappa}$ and Sec. 5.2 the spatial granularity of $\vec{V}$.

5.1. Angular optimization

Exhaustively evaluating (5), requires the value of s for every $\vec{n}$ within a cone of opening γ, to select the maximizing $\vec{\kappa}$.

To reduce the number of evaluated $\vec{n}$'s at each iteration i: (a) the cone is canonically sampled and the optimizing direction $\vec{\kappa}_i$ is selected amongst the sampled directions, (b) the sampling gets exponentially denser, but (c) only the

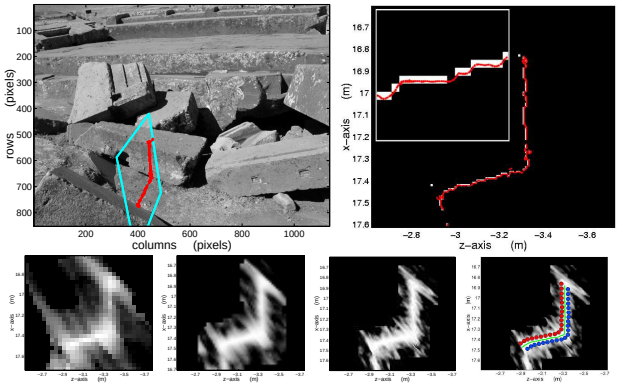

Fig. 2. Top: image from a $\approx 40\,cm$ binocular pair and SLMs along a slice of s detected with discrete (white pixels) and sub-voxel precision (red dots). Local maxima are back projected on the left image. Bottom: $s(p)$ for 3 scales; left is coarsest. At 1^{st} iteration: $\alpha_0 = 8cm$, voxel = $(4cm)^3$, $r = 21$, $\sigma_0 = 5$. Final graph illustrates the placement of pivot points and interpolating surface.

samples within the opening of cone around $\vec{\kappa}_{i-1}$ are evaluated; at the next iteration, $\gamma_{i+1} = \gamma_i/\delta$. Iterations start from the frontoparallel $\vec{\kappa}_1$ and end when γ_i falls below a precision threshold τ_γ.

Theoretically, a speedup of δ^3/γ is expected, which corresponds to ≈ 20 for $\gamma_1 = 60°, \delta = 3$ and $\tau_\gamma = 1°$. In practice, ≈ 100 is obtained, since iterations end if all samples are less than threshold. In the experiments, the mean and variance of the differences to the exhaustive search were $\approx 2°$ and $\approx 5°$, respectively.

5.2. Spatial optimization

To reduce the number of evaluated $\vec{p}$s, computation is iteratively focused at the volume neighborhoods of SLMs, based on a scale-space treatment of the input images.

At each iteration (i), $\alpha_i = \alpha_0/2^i$ and $I_{1,2}$ are convolved with a Gaussian of $\sigma_i = \sigma_0/2^i$. Also, voxel volume is reduced by $1/2^3$ and correlation is computed only at the neighborhoods of SLMs of the previous iteration.

The effect of these modulations is that at initial scales correspondences are evaluated for coarse-scale texture and successively utilize more image detail. Their purpose is to efficiently compare w_1 and w_2 at coarse scales. At these scales, the projections of points on $\mathcal{S}$ are sparse and, thus, even a minute calibration error causes significant miscorrespondence of the projections. Smoothing, in effect, decreases image resolution and, thus, more correspondences are established at coarse scales. In Fig. 2, the method is demonstrated.

6. RESULTS AND CONCLUSIONS

The methods presented in this paper, were compiled in a multiview stereo algorithm, which is presented along with results and future goals.

Algorithm At each scale, $\vec{V}$ is first evaluated by sweeping and SLMs are detected. Only at their neighborhoods, SLM detection is twice refined. First, by re-detecting SLMs by orientation search and, second by filtering $\vec{V}$ as in Sec. 4.2. At the next scale, the evaluated voxels will be restricted to the neighborhoods of the detected SLMs (see Sec. 5.2). The process is shown for one scale in Fig. 1; in cases (c) and (d), $\vec{V}$ was computed only at SLM neighborhoods.

If > 1 views exist, at the end of the iteration the space-carving rule [25] is applied to detect all empty voxels. At the next scale, each view also excludes from computation all the voxels that are known to be empty, due to any view. At the last scale the result is interpolated by the method of Sec. 4.3 and a mesh is outputted. Combining information from multiple (binocular or trinocular) views, includes selecting the maximum scoring view for each voxel to create a combined $\vec{V}$ [9].

Results The above algorithm was applied to the example of Fig. 3, for both semi-automatic calibration [26] and self-calibration (see Sec. 3). The results shows that the proposed approach is robust to the minutes inaccuracies of self-calibration. Speedup values $\approx 10^2$ and ≈ 6 compared to the exhaustive search for the finest voxel size, were obtained for the two methods of Sec. 5.1 and 5.2 respectively. For reconstructing the volume of the example, $86sec$ were required on a Pentium $3.2GHz$. The results are compared with the cases of detecting SLMs by sweeping and orientation-search. A wide outdoors area reconstruction is also shown, demonstrating the mutliview expansion of the algorithm. Finally, Fig. 4 compares the two calibration methods to show that self-calibration results sufficiently approximate the "ground truth".

Discussion For one view, each step of the algorithm can be executed in massively parallel since computation is independent per voxel [27]. The reconstruction volume is partitioned as to the number of CPUs available. The need for neighboring information (e.g. in the SLM detection) is fulfilled by overlapping the partitions by a few voxels. However, the algorithm's execution time is determined on the contents of the reconstructed volume. In multiview cases, where visibility is accounted for, "inter-view" communication is required. To optimize speedup it is required that CPUs complete their previous task (optimization) simultaneously. Thus it is required that computational load is equally balanced among CPUs. Definitely, our next experiment in this direction would be minutely partition $\vec{V}$ and adjust the supply of work towards the CPUs according to estimated completion time.

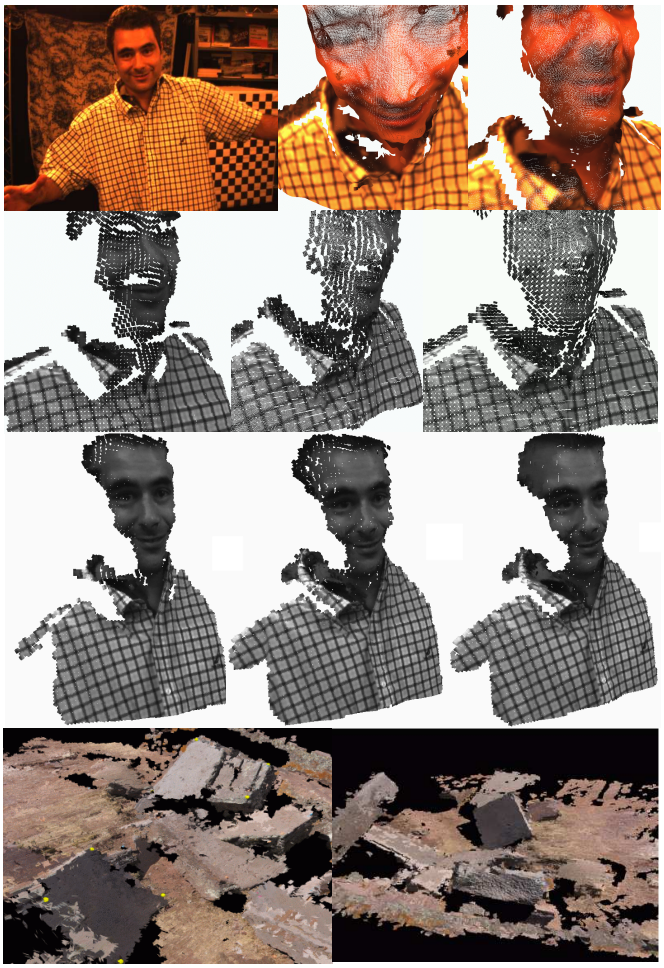

Fig. 3. Results. Top: image from a binocular pair and final reconstruction of $1m^3$ (all parameters as in Fig. 1). 2^{nd} row, left to right: detection of SLMs by sweeping, optimization search, and as in Sec. 4.2. 3^{rd} row, left to right: corresponding reconstructions as 2^{nd} row but for self-calibrated cameras. Bottom: multiview reconstruction of the scene of Fig. 2 (same parameters).

$$\begin{bmatrix} 818.06 & 1.25 & 374.75 \\ 0 & 814.35 & 248.53 \\ 0 & 0 & 1.00 \end{bmatrix} \quad \begin{bmatrix} 804.19 & 0 & 320 \\ 0 & 804.19 & 240 \\ 0 & 0 & 1 \end{bmatrix}$$

Fig. 4. Camera matrices for automatic (left) and self-calibration (right).

References

[1] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *IJCV*, vol. 47, no. 1-2-3, pp. 7–42, 2002.

[2] K. M. Cheung, T. Kanade, J. Y. Bouguet, and M. Holler, "A real time system for robust 3d voxel reconstruction of human motions," in *Proc. IEEE CVPR*, 2000, vol. 2, pp. 714 – 720.

[3] V. Kolmogorov and R. Zabih, "Multi-camera scene reconstruction via graph cuts," in *Proc. ECCV*, 2002, vol. 1, pp. 379–393.

[4] K. N. Kutulakos and S. M. Seitz, "A theory of shape by space carving," *IJCV*, vol. 38, no. 3, pp. 197–216, 2000.

[5] A. Bowen, A. Mullins, R. Wilson, and N. Rajpoot, "Light field reconstruction using a planar patch model.," in *SCIA*, 2005, pp. 85–94.

[6] R. Carceroni and K. Kutulakos, "Multi-View scene capture by surfel sampling: From video streams to Non-Rigid 3D motion, shape & reflectance," *IJCV*, vol. 49, no. 2-3, pp. 175–214, 2002.

[7] O. Faugeras and R. Keriven, "Complete dense stereovision using level set methods," in *Proc. ECCV 98*, 1998, vol. 1, pp. 379–393.

[8] A. S. Ogale and Y. Aloimonos, "Stereo correspondence with slanted surfaces: critical implications of horizontal slant," in *Proc. CVPR04*, 2004, vol. 1, pp. 568–573.

[9] X. Zabulis and K. Daniilidis, "Multi-camera reconstruction based on surface normal estimation and best viewpoint selection," in *Proc. of IEEE 3DPVT*, 2004, pp. 733–40.

[10] S. Paris, F. Sillion, and L. Qu, "A surface reconstruction method using global graph cut optimization," in *ACCV*, 2004.

[11] K. Junhwan, V. Kolmogorov, and R. Zabih, "Visual correspondence using energy minimization and mutual information," in *Proc. ICCV*, 2003, vol. 2, pp. 1033–1040.

[12] R. T. Collins, "A space-sweep approach to true multi-image matching," in *CVPR96*, 1996, pp. 358–363.

[13] J. Mulligan, X. Zabulis, N. Kelshikar, and K. Daniilidis, "Stereo-based environment scanning for immersive telepresence.," *IEEE CSVT.*, vol. 14, no. 3, pp. 304–20, 1999.

[14] Y. Ohta and T. Kanade, "Stereo by intra- and inter-scanline search using dynamic programming," *PAMI*, vol. 7, no. 2, pp. 139–154, 1985.

[15] J. Montagnat, H. Delignette, and Ayache N., "A review of deformable surfaces: topology, geometry and deformation," *Image and Vision Computing*, vol. 19, pp. 1023–1040, 2001.

[16] G. Turk and J. F. O'Brien, "Modelling with implicit surfaces that interpolate," *ACM Trans. on Graphics*, vol. 21, no. 4, pp. 855–873, 2002.

[17] C. Zeller and O. Faugeras, "Camera self-calibration from video sequences: the kruppa equations revisited," Tech. Rep. 2793, INRIA, Feb. 2005.

[18] P. R. S. Mendonca and Cipolla R., "A simple technique for self-calibration," in *CVPR*, 1999, pp. 500–505.

[19] P. R. S. Mendonca, *Multiview geometry: profiles and self-calibration*, Ph.D. thesis, University of Cambridge, 2001.

[20] M. Pollefeys and L. Van Gool, "A stratified approach to self-calibration," in *CVPR*, 1997, pp. 407–412.

[21] T. S. Huang and O. D. Faugeras, "Some properties of the e matrix in two-view motion estimation," *PAMI*, vol. 11, no. 12, pp. 1310–1312, 1989.

[22] P. Sturm, "On focal length calibration from two views," in *CVPR*, 2001, pp. 145–150.

[23] S. Bougnoux, "From projective to euclidean space under any practical situation, a criticism of self-calibration," in *ICCV*, 1998, pp. 790–796.

[24] W. Lorensen and H. Cline, "Marching cubes: A high resolution 3d surface construction algorithm," *CG*, vol. 21, no. 4, pp. 169–169, 1987.

[25] K. N. Kutulakos and S. M. Seitz, "A theory of shape by space carving," *IJCV*, vol. 38, no. 3, pp. 197–216, 2000.

[26] J. P. Barreto, K. Daniilidis, N. Kelshikar, R. Molana, and X. Zabulis, "Easycal camera calibration toolbox," Tech. Rep., Univ. of Pennsylvania, 2003, http://www.cis.upenn.edu/teleimmersion/research/downloads/EasyCal/.

[27] N. Kelshikar, Zabulis X., K. Daniilidis, V. Sawant, Sinha S., T. Sparks, S. Larsen, H. Towles, K. Mayer-Patel, H. Fuchs, J. Urbanic, K. Benninger, R. Reddy, and G. Huntoon, "Real-time terascale implementation of tele-immersion.," in *ICCS*, 2003, pp. 33–42.