# INCREASING THE ACCURACY OF THE SPACE-SWEEPING APPROACH TO STEREO RECONSTRUCTION, USING SPHERICAL BACKPROJECTION SURFACES.

*Xenophon Zabulis*[1] *and Georgios Kordelas*[1] *and Karsten Mueller*[2] *and Aljoscha Smolic*[2]

[1]Informatics and Telematics Institute, Thessaloniki, Greece

[2]Fraunhofer Institute for Telecommunications, Heinrich-Hertz-Institut, Germany

## ABSTRACT

In this paper, interest is focused on the accurate and time-efficient stereo reconstruction, for the purpose of generating 3D animated scenes from multiple synchronized videos. The plane-sweeping approach is reviewed as relevant to the goal of time-efficiency, since its execution can be optimized on a GPU. A method compatible for optimization on the GPU is proposed as a more accurate alternative to plane sweeping and to the derived visibility computation. The method is compared to plane sweeping as to its accuracy, by evaluating the backprojected 3D model against independent views and using n-fold cross validation to estimate the Peak Signal to Noise Ratio (PSNR). Finally, the method's output is casted integratable with multicamera stereo reconstruction frameworks.

## 1. INTRODUCTION

The goal of 3D television demands the ability to process large amounts of multiview video in order to extract surface reconstructions that will be stereoscopically shown. To meet this demand, the acquired images are required to be processed with time efficiency, if not in real time. In addition, multiview data are required to be processed in combination, in order to improve the accuracy of the delivered reconstructions.

In many recent multiview methods for stereo reconstruction, the acquired images are backprojected on a hypothetical surface. Ideally, if this surface coincides to the imaged one, the backprojected images should be the identical since they collineate from the same world points. In contrast, if there exists no such coincidence the backprojected image regions collineate from (or "see") different regions of the imaged surface and are, usually, dissimilar. By (a) "moving" this hypothetical surface, and (b) associating a similarity score to the position of this surface, the imaged surfaces are recovered as local similarity maxima. The most common approach is the translation (sweep) of a plane along the frontal direction.

Interest in sweeping methods was reinforced due to the ability of GPUs to optimize warping and convolution, which plane sweeping is based upon. Although not particularly accurate (see Sec. 2) the approach provides of an adequate estimate of the imaged scene. This estimate can be valuable for multiview algorithms, since it restricts the search space without necessarily imposing early decisions ($> 1$ local maxima along a line of sight may be considered).

In this paper, the geometry of sweeping is revisited and a spherical parameterization of the sweeping surface is proposed and evaluated. This approach has equal applicability to GPU optimization, since texture mapping onto spheres is supported by the corresponding hardware. The main advantage of spherical parameterization emerges when processing the periphery of the acquired images. There, a part of a frontoparallel plane is imaged obliquely and, thus, in less pixels than a part of the same plane, but imaged at the center of the image. Through comparative experiments it is concluded that the above phenomenon has a reducing effect in the accuracy of the obtained reconstruction. In the proposed approach, this effect is absent, because the corresponding image area is equal for all eccentricities of the Field of View (FOV).

The remainder of this paper is organized as follows. In Sec. 2, relevant work is reviewed. In Sec. 3, the proposed parameterization is theoretically justified and formulated. Then in Sec. 4, it is evaluated comparatively to the planar parameterization. Finally, Sec. 5 concludes and discusses the integration of the proposed approach in a multiview framework.

## 2. RELATED WORK

Stereo reconstruction based on sweeping methods has been carried out using two types of similarity measures. Photoconsistency [1, 2] and texture similarity (typically implemented using the SAD, SSD, NCC, or MNCC operators) [3, 4, 5, 6, 7]. The difficulty in using the photoconsistency similarity metric is that it requires the radiometric calibration of the cameras, which is difficult to achieve and retain. Although that in this paper texture similarity is utilized, the argument concerning the number of pixels, within which a unit area of the backprojection surface is imaged, applies to the photoconsistency similarity metric as well. Finally, some stereo reconstruction methods that utilize feature matching are also

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

based on plane sweeping, including the first formulation of this approach algorithm [8, 9, 10]. By the same argument to photoconsistency, the proposed improvement in the shape of the sweeping surface also applies to the above methods.

In this paper, besides proposing a new shape for the sweeping surface, it also is argued that this surface, or more specifically the parameterization of points on it, should follow an projective expansion. This expansion has been sometimes overlooked the literature as most papers define sweeping as just a translation of the plane in the frontoparallel direction (e.g. [8, 9, 10, 1, 4, 5, 6]). In a slightly different context than that of stereo reconstruction, however, several view synthesis approaches [11, 12] utilize an expanding plane to generate novel views, but reconstruction is not explicit. Instead, a visual hull view synthesis in outputted without necessarily estimating the 3D coordinates of the imaged points . In [13], depth values are computed, but using a photoconsistency approach suitable for a multicamera ($> 2$) setup.

The accuracy of sweeping methods is limited compared to those that account for surface orientation [14, 15, 16] and which provide preciser results, but at an increased computational cost. In Sec. 5 the proposed parameterization is integrated with [16] for acceleration of the later method. Despite its reduced accuracy, plane sweeping provides of a time efficient method for stereo reconstruction. This time efficiency is based on implementations compatible with the GPU hardware [11, 13, 6]. Due to its single instruction, multiple data architecture the GPU executes the warping and convolution operations, by computing each pixel of the result in parallel.

## 3. SPHERE SWEEPING

The proposed method sweeps an expanding spherical sector from the cyclopean eye of a binocular pair, by backprojecting the acquired images onto it. At each position, these backprojections are locally compared to associate a similarity score for, virtually, every point within the sweeped space. Surfaces are recovered by assuming the following: the locus of the maximum score on a line along the sweeping direction is the point of intersection of this line with the imaged surface.

The proposed approach introduces two modifications to plane sweeping: (1) the backprojection plane is substituted by a spherical sector and (2) the search for local similarity maxima is performed along expanding spherical sectors (as opposed to regular voxels) and along the direction of sight (as opposed to the direction of sweep). Using a spherical backprojection surface, casts vector $\vec{v}$ from the optical center perpendicular to this surface, for any eccentricity $\epsilon$ of the FOV. In contrast a planar surface becomes increasingly slanted relatively to $\vec{v}$ as $\epsilon$ moves to the periphery of the image.

In the paragraph below, it is shown that a spherical backprojection surface maximizes the number of image pixels subtended by a unit area on this surface. Due to this maximization, backprojections contain a denser sampling of the imaged surface. In contrast, when using less pixels their similarity score tends to be biased towards spuriously high values. The reason is that variance, which is effectively a normalizing factor in similarity measures, tends to be an increasing function of the number of intensity differences that are considered.

When a unit surface is oriented perpendicular to $\vec{v}$ its image area is maximized. To show the above, an area on the backprojection surface corresponding to an image similarity kernel is assumed as locally planar. As shown in Fig. 1 (left), the subtended visual angle of this area is maximized at the perpendicular posture $CD$. In any other posture (e.g. $AB$ for plane sweeping), this angle is smaller since the image area subtended is decreased by a factor of $\cos(CpA)$ in *both* tilt and slant dimensions.

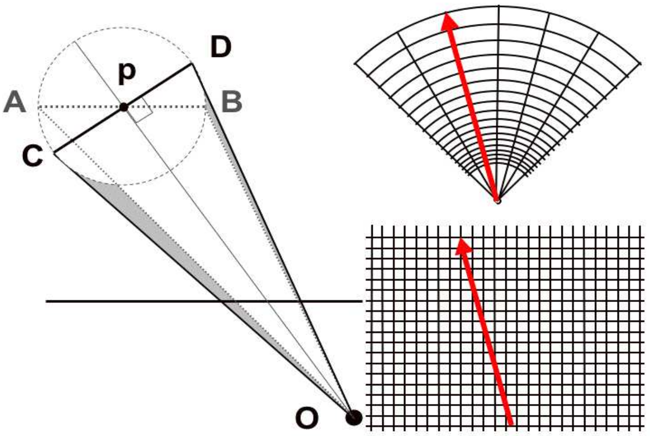

**Fig. 1**. Flatland illustrations of the geometry of sphere sweeping. *Left:* The subtended visual angle of a unit area perpendicular to the line of sight (from the projection center $O$) is greater than that of the corresponding area on a frontoparallel plane, or any other plane that $p$ may occur on. *Right:* Illustrations of the sector (top) and voxel (bottom) based volume tesselations. Visibility is naturally expressed in the first representation, whereas in the second, traversing voxels obliquely is required for its computation.

Parameterizing the tessellation of the reconstruction volume into sectors instead of voxels, provides of a useful surface parameterization because of two reasons. First, because the data required to compute visibility are already structured with respect to visibility from the optical center. Application, then, of visibility rules becomes more accurate because the oblique traversal of a regular voxel space, which leads to discretization artifacts, is avoided. Second, because the fact that the spatial granularity of surface discretization is a function of distance and image resolution. Therefore, at greater distances less representational capacity is required to represent the imaged surface, but still at the same detail.

**Method formulation** Let a series of concentric and expanding spherical sectors $S_i$ at corresponding distances $d_i$ from the cyclopean eye ($C$). Their openings $\mu$, $\lambda$ in the horizontal and vertical direction, respectively, are matched to the horizontal and vertical FOVs of the cameras and tessellated by an angular step of $c$ . Parameterization variables $\psi$ and $\omega$ are determined as $\psi \in \{c \cdot i - \mu;\ i = 0,1,2,\ldots,2\mu/c\}$ and $\omega \in \{c \cdot j - \lambda;\ j = 0,1,2,\ldots,2\lambda/c\}$ and $[\mu/c] = \mu/c$,$[\lambda/c] = \lambda/c$. Angle $\psi$ varies on the $xz$ and $\omega$ on the $yz$ plane. For both, value 0 corresponds to the $zz'$ axis.

To generate sectors $S_i$, a corresponding sector $S_0$ is first defined on a unit sphere centered at $O = [0\,0\,0]^T$. A point $p = [x\,y\,z]^T$ on $S_0$ is given by: $x = sin(\psi)$, $y = cos(\psi)sin(\omega)$, $z = cos(\psi)cos(\omega)$, Its corresponding point $p'$ on $S_i$ is then:

$$p' = d_i \left[R_z(-\theta)R_y(-\phi)p + C\right], \qquad (1)$$

where $R_y$ and $R_z$ are rotation matrices for rotations about the $yy'$ and $zz'$ axes, $\vec{v_1}$ and $\vec{v_2}$ are unit vectors on the principal axes of the cameras, $\vec{v} = (\vec{v_1} + \vec{v_2})/2$, and $\theta$ (longitude), $\phi$ (colatitude) $\vec{v}$'s spherical coordinates. Computational power is conserved, without reducing the granularity of the reconstructed depths, when parameterizing $d_i$ on a disparity basis [17]: $d_i = d_0 + \beta^i$, $i = 1,2,\ldots i_N$, where $d_0$ and $i_N$ define the sweeping interval and $\beta$ is modulated so that the farthest distance is imaged in sufficient resolution.

The rest of the sweeping procedure is mostly conventional and thus, overviewed. For each $S_i$, the stereo images ($\geq 2$) are sampled at the projections $S_i$'s points on the acquired images, thus forming two $(2\mu/c \times 2\lambda/c)$ backprojection images. These images are locally compared using a similarity metric (e.g. SSD, SAD, NCC) whose implementation in most cases can be optimally performed in GPU as a combination image difference and convolution (e.g. [7, 18]). The resulting values are attached on a sector-interpretable grid (Fig. 1(right)), but whose data are structured in memory in a conventional 3D matrix. Memory is conserved similarly to [8], where a buffer that stores only the correlation result for each depth is utilized. The difference of the proposed approach is that our buffer stores the correlation result of both the previous and the next depth, in order to determine if the maximum is local.

## 4. RESULTS

In the experiments, three sweeping methods were compared: sweeping by (A) plane translation, (B) plane expansion, (C) sector expansion. Method B is similar to C differing only in the shape of the sweeping surface. Methods were compared on the basis of their reconstruction results. Reconstruction assumed that the location of the imaged surface coincides with the locus of the strongest local similarity maximum along a ray of sweep. The results were then benchmarked by two metrics: (i) as the quantity of reconstructed surface (# of reconstructed points), and (ii) the proportion of this quantity that correctly reconstructs the imaged surfaces. Evaluation of (ii) involved comparing the reconstruction result to an independently acquired $3^{rd}$ image as explained forward. Next, experimental results are shown and discussed.

In the evaluation provided, the proposed method was tested for the most basic case of stereoscopy, that of a binocular pair. The scenes were reconstructed from 2 calibrated images, after compensating for radial distortion, and evaluated against a $3^{rd}$ independent view. The evaluation was based on a 2D error measure between the $3^{rd}$ view and the reconstruction, which is backprojected onto the $3^{rd}$ cameras plane. A conventional 2D pixel-based error measure (PSNR) similar to [3] but which differs as follows.

**Evaluation method** To obtain an overall measure for a 3D reconstruction approach, the individual PSNR measures are combined by n-fold cross validation, a statistical measure, which is often used in classification processes to determine the predictability of a data set, if a certain amount of test data is missing. Starting from a total dataset of $M$, the dataset is split into n subsets $M_i$, $i = 1,2,\ldots,n$. The common procedure in n-fold cross validation is than to define the training sets $T_i = M - M_i$, predict the missing subset $M_i$ and evaluate the error between prediction and subset for all $i$. This process is also applied vice versa by using only a certain subset $M_i$ and predicting all other subsets $\{M_k \,|\, k = 1,2,\ldots,n,\, k \neq i\}$.

**Experiment and discussion** To fairly compare the methods, the same similarity thresholds (correlation = .7), FOVs ($\frac{\pi}{4}$, $\frac{\pi}{4}$) and a $d_i$ tessellation ($2mm$, regular) were utilized. The density of sample points on the different backprojection surfaces was set to be by average equal to frequency of the regular sampling of (A) ($2mm$). The baseline of the binocular pairs was $26cm$ and the $3^{rd}$ camera (for the evaluation) was in the midpoint and $\approx 13cm$ above the baseline. Images were $480 \times 640$ pixels and target surfaces occurred from $1m$ to $3m$ from the cameras.

The methods were applied at the frames of a multiview sequence and the corresponding scenes reconstructed (Fig. 2a). In Fig. 2, slices along depth that were extracted from these reconstructions at different eccentricities are compared. Almost no effect between the three methods was observed in the reconstructions, when obtained from the center of images (Fig. 2b) and, as expected, most differences were observed at the periphery of images. Methods B and C differed the most when their action was compared in the part of the reconstruction obtained from the periphery of images (Fig. 2cde); quantitatively, method C provided about $\approx 15\%$ more matches as shown in Fig. 3. The evaluation results for reconstruction of the whole scene of the examples above are reported in Fig. 3. For the example that compared the performance in a central image eccentricity B slightly outperformed C. However, in the rest of the cases where the more peripheral part of the image are considered, C clearly outperforms the other methods.

## 5. CONCLUSION

In this paper, a method that proposes the use of a spherical backprojection surface in space sweeping stereo reconstruction methods was introduced. It is concluded (see Sec.4) that

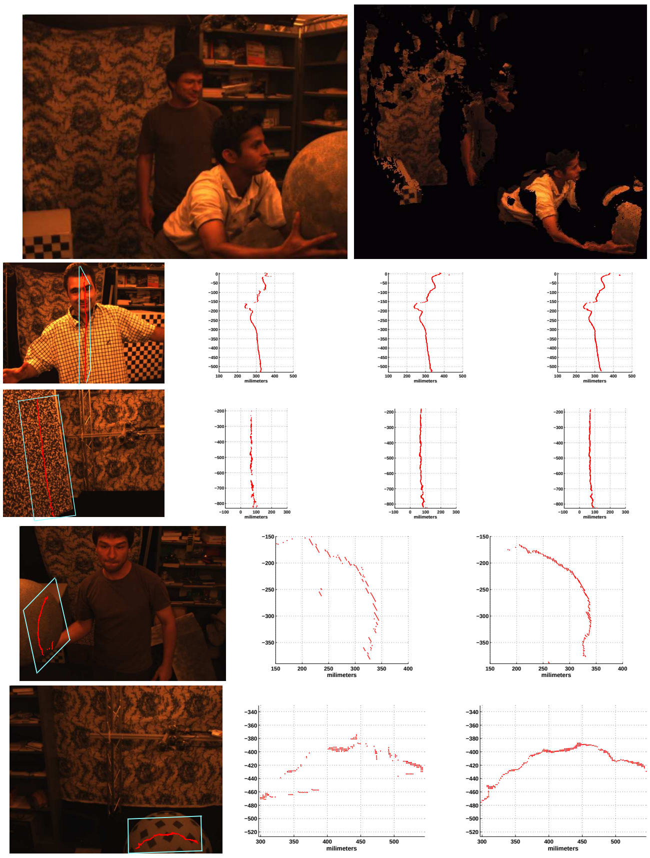

**Fig. 2**. Comparison of results; an image from a binocular pair and sections of the obtained reconstructions. Corresponding sections and reconstructions are superimposed on each image. *Top to bottom:* (a) scene image and corresponding sphere sweeping reconstruction (b,c) comparison of methods A, B, and C and (d,e) comparison of methods A and C.

the spherical parameterization of the sweeping surface provides both a quantitative and qualitative improvement in the reconstruction of surfaces. Other than the shape of the sweeping surface, the implementation of the method is similar to that of plane sweeping. Thus, since GPUs can warp images onto spheres in the same way that the warp onto planes, the proposed parameterization is applicable to GPU-optimized implementations of the sweeping algorithm.

A critique against the proposed method could concern the irregular parameterization of the reconstruction volume into sectors. Indeed, in multiview stereo reconstruction cases the knots of the sector grids for each binocular pair do not coincide, causing the absence of a common reference frame to combine the views. To cope with this problem, the similarity scores are interpolated to a regular voxel grid (Fig. 4).

The proposed method is utilized as a first step in multiview stereo to reduce computational cost. After reconstruction with the proposed method, voxels corresponding to the reconstructed surfaces are considered for preciser surface localization and multiview visibility reasoning [16].

|  | A (dB) | A (#) | B (dB) | B (#) | C (dB) | C (#) |
|---|---|---|---|---|---|---|
| 1 | 28.36 | 35797 | 31.26 | 36705 | 30.71 | 38074 |
| 2 | 35.19 | 9851 | 35.01 | 10701 | 36.40 | 12957 |
| 3 | 42.80 | 10874 | 45.44 | 11822 | 45.68 | 12322 |
| 4 | 43.29 | 10098 | 43.28 | 12327 | 44.28 | 12871 |

**Fig. 3**. PSNR evaluation ($dB$) and number of reconstructed points (#): results for the whole reconstruction of the examples in Fig 2.

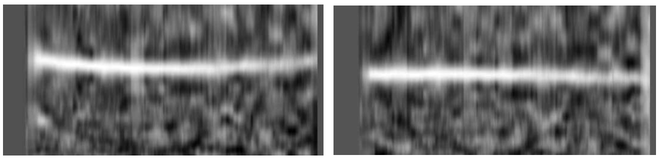

**Fig. 4**. Similarity scores across the section of the reconstruction volume of Fig. 2b. *Left:* pixel locations correspond to knots on a sector grid. it Right: the same values interpolated to a regular grid.

## References

[1] K. N. Kutulakos and S. M. Seitz, "A theory of shape by space carving," *IJCV*, vol. 38, no. 3, pp. 199–218, 2000.

[2] S. M. Seitz and C. R. Dyer, "Photorealistic scene reconstruction by voxel coloring," *IJCV*, vol. 35, no. 2, pp. 151–173, 1999.

[3] R. Szeliski, "Prediction error as a quality metric for motion and stereo," in *ICCV*, 1999, vol. 2, pp. 781–788.

[4] C. Zhang and T. Chen, "A self-reconfigurable camera array," in *Eurographics Symposium on Rendering*, 2004.

[5] T. Werner, F. Schaffalitzky, and A. Zisserman, "Automated architecture reconstruction from close-range photogrammetry," in *CIPA*, 2001.

[6] C. Zach, A. Klaus, B. Reitinger, and K. Karner, "Optimized stereo reconstruction using 3D graphics hardware," *Workshop of Vision, Modelling, and Visualization*, pp. 119–126, 2003.

[7] I. Geys, T. P. Koninckx, and L. J. Van Gool, "Fast interpolated cameras by combining a gpu based plane sweep with a max-flow regularisation algorithm," in *3DPVT*, 2004, pp. 534–541.

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

[9] J. Bauer, K. Karner, and K. Schindler, "Plane parameter estimation by edge set matching," in *26th Workshop of the Austrian Association for Pattern Recognition*, 2002, pp. 29–36.

[10] C. Zach, A. Klaus, J. Bauer, K. Karner, and M. Grabner, "Modelling and visualizing the cultural data set of Graz," in *VAST*, 2001.

[11] M. Li, M. Magnor, and H. P. Seidel, "Hardware-accelerated rendering of photo hulls," *Eurographics*, vol. 23, no. 3, 2004.

[12] V. Nozick, S. Michelin, and D. Arqus, "Image-base rendering using plane-sweeping modelisation," in *IAPR MVA*, 2005, pp. 468–471.

[13] R. Yang, G. Welch, and G. Bishop, "Real-time consensus-based scene reconstruction using commodity graphics hardware," in *Pacific Graphics*, 2002.

[14] 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.

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

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

[17] J. X. Chai, X. Tog, S. C. Chan, and H. Y. Shum, "Plenoptic sampling," in *SIGGRAPH*, 2000, pp. 307–318.

[18] C. Zach, K. Karner, and H. Bischof, "Hierarchical disparity estimation with programmable 3D hardware," in *WSCG*, 2004, pp. 275–282.