MODULATING THE SIZE OF BACKPROJECTION SURFACE PATCHES, IN VOLUMETRIC STEREO, FOR INCREASING RECONSTRUCTION ACCURACY AND ROBUSTNESS

Xenophon Zabulis and Georgios D. Floros

Informatics and Telematics Institute, Centre for Research and Technology Hellas
xenophon@iti.gr

ABSTRACT

This paper concerns volumetric stereo methods, which compare the backprojections of the acquired images onto a hypothetical surface patch in order to reconstruct the imaged surfaces. In particular, it introduces a size-modulation of this patch so that its projection area in the acquired images is invariant to distance and rotation. It is shown, and explained why, that performing this modulation results in superior accuracy of the volumetric reconstruction than retaining the patch size constant, as it has been to date practiced. The proposed extension to the hypothetical patch operator is compatible with the existing volumetric approaches to stereo.

Index Terms— stereo, volumetric stereo, reconstruction, backprojection, surfel

1. INTRODUCTION

Stereo-correspondences among a pair of images have been traditionally established by the similarity-matching of the image neighborhoods of the potentially matching points (e.g. [1]). More recently [2, 3, 4, 5], stereo approaches have introduced a *volumetric* and more robust [6] way of performing this similarity-matching, by comparing the backprojections of the acquired images on hypothetical surfaces that are tangent to the imaged ones. This paper is related to the latter approach and its contribution consists in the improvement of the, above, volumetric similarity-matching approach in terms of accuracy and robustness.

Contrary to traditional stereo where image neighborhoods are compared, volumetric stereo compares the backprojections $w_i(\vec{p},\vec{n})$ of the acquired pair of images ($I_{1,2}$) on some hypothetical planar patch, or surfel, $\mathcal{S}$. When $\mathcal{S}$ is tangent at a world surface, $w_i(\vec{p},\vec{n})$ are identities of the - same - surface pattern and should, thus, exhibit optimal similarity. Otherwise, $w_{1,2}$ are dissimilar because they are collineations from different surface regions. Therefore, volumetric approaches attempt to detect the locations and orientations that optimize the similarity of $w_{1,2}$ in order to reconstruct surfaces.

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

To perform this detection, the similarity of $w_{1,2}$ is evaluated at potential locations and orientations of $\mathcal{S}$ and the locations at which similarity is optimized are considered as occurring on the imaged surface. Function $\vec{V}(\vec{p}) = s(\vec{p})\vec{\kappa}(\vec{p})$, is evaluated as:

$$s(\vec{p}) = \max_{\vec{n}} \left(sim(w_1(\vec{p},\vec{n}), w_2(\vec{p},\vec{n})) \right), \qquad (1)$$

$$\vec{\kappa}(\vec{p}) = \arg\max_{\vec{n}} \left(sim(w_1(\vec{p},\vec{n}), w_2(\vec{p},\vec{n})) \right), \qquad (2)$$

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 [7], etc or the photoconsistency metric [8]; henceforth MNCC is utilized in all examples. To evaluate sim, a $r \times r$ lattice of points is assumed on $\mathcal{S}$. The image formation rule for $w_{1,2}$ for a square $\mathcal{S}$ of size $\alpha \times \alpha$ is:

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

where P_i is the projection matrix of I_i, $R(\vec{n}) \cdot [0\,0\,1]^T = \vec{n}$, and x', $y' \in [-\frac{\alpha}{2}, \frac{\alpha}{2}]$ local coordinates on $\mathcal{S}$.

The parameterization of posture κ refers to the orientations within a cone of opening γ around orientation $\vec{c}$ and can be expressed in terms of longitude and latitude. This orientation is defined as $\vec{c} = \vec{p} - \vec{e}$ from the cyclopean eye $\vec{e}$ to $\vec{p}$, so that different eccentricities in the image are equally treated. Analytical formulas for the above parameterization of $\vec{\kappa}$ can be found in [9]. The computational cost of the optimization for a voxel is $\mathcal{O}(Nr^2)$, where N the number of postures evaluated in $\vec{\kappa}$'s optimization. In this paper, it is assumed that images portray Lambertian surfaces, which can also be locally approximated by planar patches. Extension of these concepts beyond the Lambertian domain can be found in [10].

A similarity among the volumetric methods found in the literature is that the size of $\mathcal{S}$ is independent of the relative obliqueness and distance of the patch to the cameras. In this work, a size-modulation of $\mathcal{S}$ is proposed, such that the image area subtended by $\mathcal{S}$ is invariant to its distance from the cyclopean eye and its orientation. For reasons explained in Section 3, this normalization increases the robustness of the patch operator as a cue to the stereo-reconstruction of 3D surfaces.

The remainder of this paper is organized as follows. In Section 2 related work is reviewed. The size-modulation rule, relevant theory and experiments are presented in Section 3. Finally, in Section 4, this work is placed in the existing context of volumetric approaches and conclusions are drawn.

2. RELATED WORK

Two comprehensive reviews of the literature on stereo algorithms can be found in [6, 11]. In this section, emphasis is placed upon methods that backproject and compare images on hypothetical surfaces in space. Such methods can be classified under two categories.

The first regards *space sweeping* algorithms, where a single hypothetical surface is swept along depth [12, 13, 14, 15, 16, 17]. The orientation of the sweeping surface is a priori set independently to the actual structure of the imaged scene. Typically, it coincides with the viewing direction. The backprojections of the acquired images on this surface are *locally* compared as to their visual similarity and the results are stored in a 2D similarity map. A depth-ordered stack of such similarity maps is generated and for each column along depth, the depth at which similarity is maximized is considered to signify the occurrence of the imaged surface. Algorithms of this category mainly differ in the shape of the sweeping surface [15, 13, 14] or the collineated image content [12, 18, 19].

In the second category [2, 3, 4, 5], the acquired images are backprojected on a hypothetical planar surface patch, which is assumed at each voxel of the reconstruction volume. The backprojections are then compared as to their visual similarity, instead of regular patches in the acquired images. If the voxel turns out to be occupied by some surface, the orientation that maximizes this similarity is an estimation of the orientation of the imaged surface. This category of algorithms provides of more accurate results than the first [6] and differs into that a whole range of potential orientations is considered instead of a sinlge one. Because of this optimization, the compared textures are compensated for projective distortion, thus leading to more robust matching. The reason of this robustness is that the "correct" backprojections are devoid of perspective distortion and, therefore, expected to be more visually similar than their corresponding counterparts in the acquired images. The outcome is that similarity matching becomes more robust, thus facilitating more accurate results.

In the second category, the size of the patch has not been modulated in the literature, except in [4, 5] where a coarse to fine approach was employed in the evaluation of Equations 1 and 2, but for the acceleration of the computation. The modulation in these cases was, however, was discrete and identical for any location within the reconstruction volume and, also, irrelevant to the orientation of the patch.

3. SIZE-MODULATION

In this section, the size α of $\mathcal{S}$ is studied and its modulation is proposed to increase the accuracy of the operator defined in Equation 1.

In discrete images, the number of subtended pixels by the projection of $\mathcal{S}$ is analogous to the reciprocals of distance squared and the cosines of relative tilt and slant of $\mathcal{S}$ to the cameras. For $w_{1,2}$, when α is constant the greater the obliqueness of $\mathcal{S}$ (relative to the cameras), the fewer the image pixels that the $r \times r$ image samples are obtained from. Therefore, there will always be some level of obliqueness above which the same image intensity values will be multiply sampled. In this case, as obliqueness and/or distance increase, the population of these intensities will tend to exhibit reduced variance; because they are being sampled from decreasingly fewer pixels. Thus, a bias in favor of greater slants and distances is predicted. Mathematically, because variance occurs in the denominator of the correlation function. Intuitively, because fewer image area supports now the similarity matching of backprojections on $\mathcal{S}$, and as a consequence, this matching becomes less robust to lack of resolution.

3.1. Experiments with patches of constant size

To observe the predicted phenomenon, surface orientation was estimated and compared to ground truth, using Equations 1 and 2. Two separate experiments, with synthetic and real images, were conducted. Synthetic images were utilized to stress that the discussed inaccuracy cannot be attributed to noise or calibration errors. Therefore, the inaccuracy must be an effect of the information loss, due to image discretization. Real images, with known ground truth, were utilized to observe that the combination of the discussed phenomenon with noise gives rise to even larger estimation errors.

In the first experiment (see Figure 1), a binocular image pair was synthesized to portray a square, textured and planar piece of surface. Equations 1 and 2 were then evaluated for the central point on the surface. In the similarity map, lighter pixels indicate a high similarity value and darker the opposite (henceforth, this convention is followed for all the similarity maps in this paper). The experiment was repeated for a surface pattern at coarser scale. A longitude - latitude similarity map was in each case generated, showing s for each instance of $\vec{n}$, where longitude and latitude refer to the parameterization of $\vec{p}$. In this parameterization, the pole of the hemisphere is indicated by $\vec{c}$, longitude is the azimuthal coordinate and latitude the elevation from the equator. As shown in the figure, the predicted inaccuracy was pronounced and observed, but in small magnitude due to the perfect calibration of the synthetic images.

In the second experiment, calibration inaccuracies give rise to even more misleading local maxima in the similarity map and, also, the similarity value at very oblique pos-

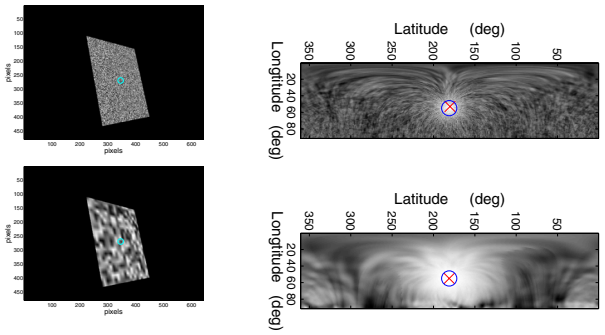

Fig. 1. Accuracy evaluation of the patch operator. Two textures rendered on a $251 \times 251\,mm$ planar surface patch, centered at $\vec{p}$ and oriented as $[001]^T$ (left). The corresponding similarity values $sim(w_1(\vec{p},\vec{n}), w_2(\vec{p},\vec{n}))$ are in a longtitude-latitude parameterization of $\vec{n}$ (right). In the maps, camera pose $\vec{c}$ is at $(0,0)$, crosses mark the maximal similarity value and circles mark ground truth. In the experiment, $\alpha = 100mm$, $r = 15$, $\gamma = 90°$. The binocular pair was $\approx 1.5\,m$ from the patch and its baseline was $26\,cm$. The, angular, discretization of the postures of $\vec{n}$ was $.5°$. The errors for the two textures were and $2.218°$ (top) and $0.069°$ (bottom). The reported errors refer to the angle between the ground truth normal and the estimated one.

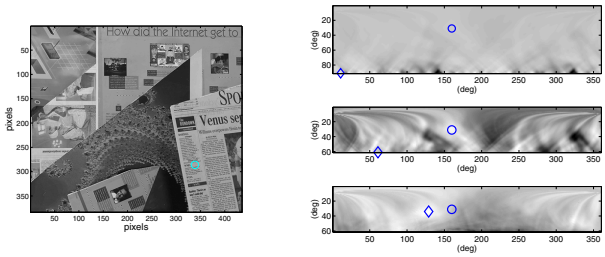

Fig. 2. Repeating the experiment of Figure 1 for the first two frames of the "Venus" Middlebury Sequence and for two different values of γ: $120°$ (top map) and $180°$ (middle map); longitude / latitude axes as in Figure 1. In the experiment, $\alpha = 250$ length units, baseline was 100 length units and $r = 151$. The surface point for which Equations 1 and 2 were evaluated is marked with a diamond. The projection of $\mathcal{S}$ subtended ≈ 50 pixels. The bottom map shows the increase in accuracy obtained by the size-normalization of $\mathcal{S}$ with respect to obliqueness (see forward in text). Mapping of similarity values to intensities is individual for each map.

tures ($> 60°$) is observed to take extreme positive or negative values. To indicate the rise of the spurious maximum at the extremes of the correlation map, the optimization was twice computed, in Figure 2: once for $\gamma = 120°$ and once for $\gamma = 180°$. In each case, the global maximum occurred at the extreme border of this map, which corresponds to a posture more oblique than ground truth.

3.2. The proposed size-modulation

The above inaccuracy can be alleviated if the size α of $\mathcal{S}$ is modulated so that the image area at which $\mathcal{S}$ is projected remains invariant, while $\mathcal{S}$ is hypothesized at different postures and distances from the cameras. In particular, the side of $\mathcal{S}$ (or diameter, for a circular $\mathcal{S}$) is varied as:

$$\alpha = \alpha_0 \left(\frac{d}{d_0} \cdot \frac{1}{\cos\omega} \right) \; ; \quad \omega = \cos^{-1}\left(\frac{\vec{v}\cdot\vec{\kappa}}{|\vec{v}|\cdot|\vec{\kappa}|} \right), \qquad (4)$$

where $\vec{v}$ is the vector from the optical center to $\vec{p}$, $d = |\vec{v}|$, ω is the angle between $\vec{v}$ and $\vec{\kappa}$ and d_0, α_0 initial parameters in units of world length. In the above equation, factor d/d_0 normalizes for changes in distance, while $(\cos\omega)^{-1}$ for changes in posture. Figures 2 and 3 show the spatial and angular effect obtained by the proposed size-modulation and compare the obtained reconstructions with those obtained without its employment. As shown in these figures, increased accuracy is obtained utilizing the proposed normalization. In the first case (orientation experiment), the surface normal is more accurately estimated for a given point on the image surface. In the second, the locations in 3D space of the imaged surface are more accurately defined.

Finally, a "side-effect" of this normalization is that the larger the distance and obliqueness of a surface, the lower the spatial frequency (not resolution) that is reconstructed at. This effect is considered as a fair tradeoff, since distant and oblique surfaces are imaged at lower frequencies, anyway.

4. CONCLUSION

In this paper, it is argued that modulating the size of the hypothetical planar patch operator in volumetric stereo method, so as the patch to project at an equal amount of image area for each locus and orientation, produces more accurate results than when retained constant. The increase in robustness of the proposed approach versus approaches that utilize a patch of constant size was confirmed through reconstruction experiments, where ground truth was known.

Volumetric stereo algorithms utilize the readings of the hypothetical planar patch operator in different ways. For example, in [3] similarity values are input into a global optimization the result of which is an isosurface that represents the reconstructed surface. In [2] besides geometrical, photometrical properties are also computed on the patch and a multi-dimensional optimization is employed to determine the occupied voxels. In [4, 5], spatially local maxima in the response of the operator are regarded as due to surface occurrence. It is argued that the proposed modulation can be directly adopted by volumetric methods, such as the above, that utilize a constant-size hypothetical patch. The reason is that the modulation is defined as reparameterization dependent only on the assumed location and orientation of the patch.

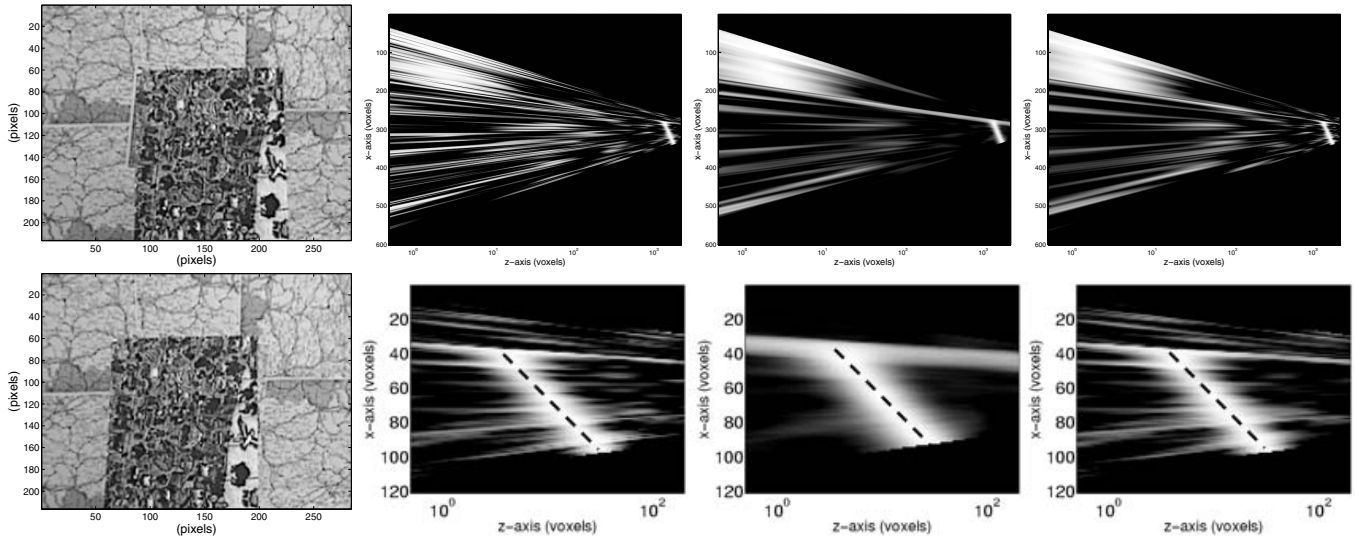

Fig. 3. Comparison of techniques. Shown is a stereo pair (left column) and three separate calculations of s across a vertical section, through the middle of the foreground surface. The bottom figures are zoom-in detail on the part that corresponds to the foreground and z-axes (horizontal in maps) are logarithmic. In the 2^{nd} column, a fine α was used, hence the noisy response at the background. Using a coarse α (3^{rd} column), yields a smoother response in greater distances, but diminishes any detail that could be observed in short range. In the 4^{th} column, α is projectively increased, thus normalizing the precision of reconstruction by the area that a pixel images at that distance. In the bottom figures, ground truth is marked with a dashed line.

5. REFERENCES

[1] N. Ayache., *Artificial Vision for Mobile Robots: Stereo Vision and Multisensory Perception*, MIT Press, Cambridge MA, 1991.

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

[3] O. Faugeras and R. Keriven, "Complete dense stereovision using level set methods," in *European Conference on Computer Vision*, 1998, vol. 1, pp. 379–393.

[4] X. Zabulis and K. Daniilidis, "Multi-camera reconstruction based on surface normal estimation and best viewpoint selection," in *IEEE International Symposium on 3D Data Processing, Visualization and Transmission*, 2004, pp. 733–40.

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

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

[7] H. Moravec, "Robot rover visual navigation," *Computer Science: Artificial Intelligence*, pp. 105–108, 1980/1981.

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

[9] X. Zabulis and G. Kordelas, "Efficient, precise, and accurate utilization of the uniqueness constraint in multi-view stereo," in *IEEE International Symposium on 3D Data Processing, Visualization and Transmission*, 2006.

[10] H. Jin, S. Soatto, and A. Yezzi, "Multi-view stereo beyond lambert," in *IEEE Computer Vision and Pattern Recognition*, 2003, pp. 171–178.

[11] M. Z. Brown, D. Burschka, and G. D. Hager, "Advances in computational stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 8, pp. 993–1008, 2003.

[12] R. T. Collins, "A space-sweep approach to true multi-image matching," in *IEEE Computer Vision and Pattern Recognition*, 1996, pp. 358–363.

[13] X. Zabulis, G. Kordelas, K. Mueller, and A. Smolic, "Increasing the accuracy of the space-sweeping approach to stereo reconstruction, using spherical backprojection surfaces," in *International Conference on Image Processing*, 2006.

[14] M. Pollefeys and S. Sinha, "Iso-disparity surfaces for general stereo configurations," in *European Conference on Computer Vision*, 2004, pp. 509–520.

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

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

[17] V. Nozick, S. Michelin, and D. Arqus, "Image-base rendering using plane-sweeping modelisation," in *IAPR Conference on Machine Vision Applications*, 2005, pp. 468–471.

[18] 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.

[19] C. Zach, A. Klaus, J. Bauer, K. Karner, and M. Grabner, "Modelling and visualizing the cultural data set of Graz," in *International Symposium on Virtual Reality, Archaelogy and Cultural Heritage*, 2001.