Spatiotemporal Stereo via Spatiotemporal Quadric Element (Stequel) Matching

Mikhail Sizintsev and Richard P. Wildes
Department of Computer Science and Engineering
York University
Toronto, ON, Canada

Abstract

Spatiotemoral stereo is concerned with the recovery of the 3D structure of a dynamic scene from a temporal sequence of multiview images. This paper presents a novel method for computing temporally coherent disparity maps from a sequence of binocular images through an integrated consideration of image spacetime structure and without explicit recovery of motion. The approach is based on matching spatiotemporal quadric elements (stequels) between views, as it is shown that this matching primitive provides a natural way to encapsulate both local spatial and temporal structure for disparity estimation. Empirical evaluation with laboratory-based imagery with ground truth and more typical natural imagery shows that the approach provides considerable benefit in comparison to alternative methods for enforcing temporal coherence in disparity estimation.

1. Introduction

In a 3D dynamic environment a visual system must process image data that derives from both the temporal and spatial scene dimensions. Correspondingly, stereo and motion are two of the most widely researched areas in computer vision. Within this body of research, integrated investigation of stereo and motion has received relatively little attention. Ultimately, however, recovery of 3D scene structure must respect dynamic information to ensure that estimates are temporally consistent. Further, in situations where instantaneous multiview matching is ambiguous (e.g., weakly textured surfaces or epipolar aligned pattern structure), dynamic information has the potential to resolve correspondence by further constraining possible matches.

In response to the above observations, this paper describes a novel approach to recovering temporally coherent disparity estimates from a sequence of binocular images. The key idea is to base stereo correspondence on matching primitives that inherently encompass both the spatial and temporal dimensions of image spacetime. In particular, each temporal stream of imagery is locally represented in terms of its orientation structure, as captured by the spatiotemporal quadric (also variously refered to as the orientation tensor and covariance matrix, see, e.g., [10]). It will be shown that by basing matching on this representation, it is possible to recover temporally coherent disparity estimates, without the need to make optical or 3D flow explicit. Further, this representation allows spatial and temporal image structure to resolve otherwise ambiguous matches in a fashion consistent with both sources of information.

Early work combining stereo and motion concentrated on punctate features (e.g., edges, corners). One of the earliest attempts made use of heuristics for assigning spatial and temporal matches based on model-based reasoning [12]. A rather different early approach exploited constraints on the temporal derivative of disparity [26]. Other work matched binocular features to recover 3D estimates for temporal tracking [29]. More recent research that relies on loose coupling of stereo and motion has emphasized the recovery of 3D motion using optical flow in conjunction with multiple hypothesis disparity maps [5]. The proposed research differs from such early work in being focused on a more integrated approach to spatiotemporal processing and in its emphasis on dense reconstruction.

More recent stereo research has seen increased interest in scene recovery from multicamera (especially binocular) video as constrained by 3D models. Some work has concentrated on the recovery of surface mesh models between individual stereo pairs with tracking across time instances serving to yield temporally consistent models [16]. Other research considers multiple cameras, employs voxel carving for initial estimation and uses intensity-based matching over spatiotemporal volumes without consideration of image motion differences between different views [18]. Still other work casts stereo and motion estimation as a generic image matching problem solved variationally after backprojecting the input images onto a suitable surface [20]. Again, the present work puts emphasis on a more integrated approach to stereo and motion and in eschewing explicit surface models, which can become problematic when dealing with multiple objects and complex scenes.

Other lines of recent research have emphasized more integrated approaches to stereo and motion. Some of this work has concentrated on static scenes with variable lighting [4]. Others have focused on defining appropriate tempo-

978-1-4244-3991-1/09/$25.00 ©2009 IEEE

ral integration windows, e.g., as part of correspondence [28] or simply reinforce disparity estimates from the previous frame using optical flow [9]. Further, combined stereo and motion estimation has been formulated in terms of PDEs [25, 11] as well as MRFs [27, 15]. Still other work has used direct methods for integrated recovery of structure and egomotion [24, 17]. The proposed research shares with these efforts an emphasis on tight integration of binocular imagery with time. It is novel in basing its matching on the representation of image spacetime in terms of local spatiotemporal orientation, which provides richer image descriptions than raw image intensities.

A major tool that is employed in the proposed approach is the representation of spacetime imagery in terms of oriented spatiotemporal structure. Various research documents optical flow recovery [2], tracking [3] and grouping [7] using spatiotemporal orientation tuned filters. More specifically, previous research has used the spatiotemporal quadric to capture orientation in image spacetime, with application to motion estimation, restoration and enhancement [10]. However, it appears none has exploited spatiotemporal orientation, in general, or the spatiotemporal quadric, specifically, for stereo disparity estimation. Previous stereo work has defined binocular correspondence based on a bank of spatial filters [13]. The proposed approach also extracts its measures of orientation via application of a filter bank; however, it is significantly different in employing filters that span both the space and time domains, thereby basing matching on a fundamentally richer representation.

In the light of previous research, the main contributions of this work are as follows. (i) The spatiotemporal quadric is proposed as a matching primitive for spatiotemporal stereo. This primitive captures both local spatial and temporal structure and thereby enables matching to account for both sources of data without need to estimate optical flow or 3D motion. (ii) The geometric relationships between corresponding spatiotemporal quadrics across binocular views are derived and used to motivate a match cost. The spatiotemporal match primitives and cost are incorporated in local and global matchers. (iii) Extensive empirical evaluation of these matchers is presented. Testing encompasses quantitative evaluation on laboratory acquired binocular video with ground truth and qualitative evaluation on more naturalistic imagery. The datasets and associated ground truth are available for download [23].

2. Technical Approach

2.1. Spatiotemporal matching primitive

In dealing with temporal sequences of binocular images, it is possible to conceptualize of stereo correspondence in terms of image spacetime, which naturally encompasses both spatial and temporal characteristics of local pattern structure, see Fig. 1a. While image spacetime can be operated on directly, using intensities, consideration of local spatiotemporal orientation provides access to a richer representation. Local orientation has visual significance: orientations parallel to the image plane capture the spatial pattern of observed surfaces (*e.g.* spatial texture); whereas, orientations that extend into the temporal dimension capture dynamic aspects (*e.g.* motion). By integrating the temporal dimension into the primitive, matching will be inherently constrained to observe temporal coherence. Further, via combination of temporal and spatial structure in the descriptor, match ambiguities that might exist through consideration of only one data source have potential to be resolved.

To extract a representation of orientation from imagery, one can filter the data with oriented filters. In the current work, 3D Gaussian, second-derivative filters, G_2, and their Hilbert transforms, H_2 [8], are applied to the data with responses pointwise rectified (squared) and summed. Filtering is executed across a set of 3D orientations given by unit column vectors, $\hat{\mathbf{w}}_i$. Hence, a measure of local energy, E, is computed according to

$$E(\mathbf{x}; \hat{\mathbf{w}}_i) = [G_2(\hat{\mathbf{w}}_i) * I(\mathbf{x})]^2 + [H_2(\hat{\mathbf{w}}_i) * I(\mathbf{x})]^2, \quad (1)$$

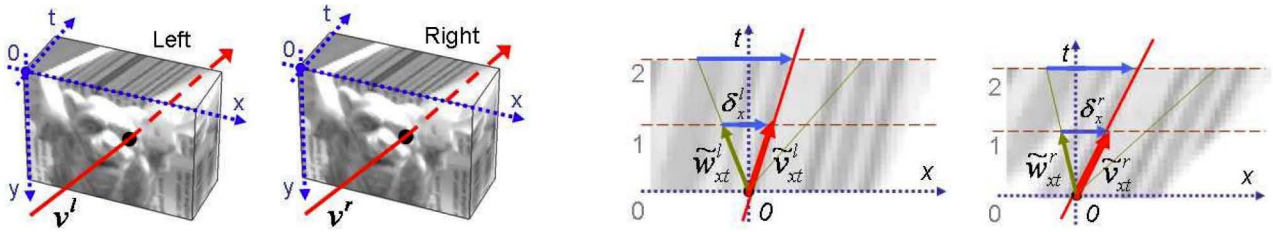
where $\mathbf{x} = (x, y, t)$ are spatiotemporal image coordinates, I is the image sequence and $*$ denotes convolution [8]. Filtering is applied separately to left and right image sequences. Here, filters are oriented along normals to icosahedron faces with antipodal directions identified, as this uniformly sample the sphere and spans 3D orientation for the employed filters. After filtering, every point in spacetime has an associated set of values that indicate how strongly oriented the local structure is along each spacetime direction.

To proceed, the individual energy measures are recast in terms of the spatiotemporal quadric. This particular representation captures local orientation as well as the variance of spacetime about that orientation. This construct captures the local shape of spacetime (*e.g.* point- vs. line- vs. plane-like) in addition to direction for a local descriptor that is richer than if (dominant) orientation alone is considered [10]. Furthermore, the quadric casts structure in terms of spacetime coordinates, $\mathbf{x} = (x, y, t)$, where it is convenient to formulate binocular match constraints. In the context of binocular matching, this quadric will be referred to as the **stequel**, *s*patio-*te*mporal *qu*adric *el*ement, Q. In particular,

$$\mathsf{Q} = \sum_i \hat{E}_i \hat{\mathbf{w}}_i \hat{\mathbf{w}}_i^\top, \quad (2)$$

where summation is across the set of filter orientations, $\hat{\mathbf{w}}_i$, and $\hat{E}_i$ is the corresponding local energy response (1) normalized such that $\sum_i \hat{E}_i(\mathbf{x}) = 1$. In constructing Q, the dyadic product, $\hat{\mathbf{w}}_i \hat{\mathbf{w}}_i^\top$, sets the local frame implied by orientation $\hat{\mathbf{w}}_i$ weighted by its response, $\hat{E}_i$, [10].

For a binocular sequence, the stequel, Q, is computed pointwise in spacetime and separately for the left and right image sequences to provide matching primitives; thus, it is



(a) (b) (c)

Figure 1. Image Spacetime. (a) Spacetime can be conceptualized as a spatiotemporal volume xyt. An instantaneous motion trajectory, v (shown in red), traces an orientation in this volume. (b) An exemplar xt slice of the spatiotemporal volume for the left view (c) The corresponding xt slice in the right view. $\widetilde{v}_{xt}^l$ and $\widetilde{v}_{xt}^r$ are the projections of the v^l and v^r onto the xt slice; w^l and w^r are arbitrary vectors (shown in green) in correspondence in xyt space and $\delta^r = \widetilde{w}^r - \widetilde{v}^r$, $\delta^l = \widetilde{w}^l - \widetilde{v}^l$ (shown in blue); $\delta^r = \mathsf{A}\delta^l$ as explained in text.

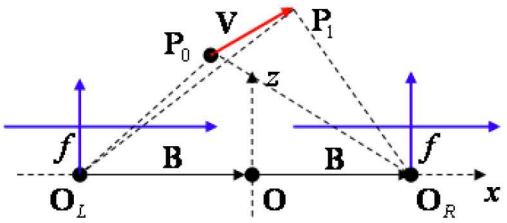


Figure 2. Stereo Geometry. A Euclidean coordinate system is set at the midpoint of the stereo baseline, $\mathbf{O}$. Cameras are rectified with a half-baseline $\mathbf{B} = [B, 0, 0]^\top$ and focal lengths f. Left and right optical centres are at $O^l = -\mathbf{B}$ and $O^r = \mathbf{B}$, resp. Point $\mathbf{P}$ undergoes an arbitrary displacement $\mathbf{V}$ from time 0 to 1.

parametrized as $\mathsf{Q}^l(\mathbf{x})$ and $\mathsf{Q}^r(\mathbf{x})$, in reference to the left and right views, resp. Significantly, the implied calculations are modest. The calculation of local energy is realized through steerable filters requiring only 3D separable convolution and pointwise nonlinearities and is thereby amenable to compact, efficient implementation [6]. Construction of Q from the filter responses requires only matrix summation, as specified in (2).

2.2. Spatiotemporal epipolar constraint

In establishing correspondence between binocular sequences, it is incorrect simply to seek the most similar stequels, as local spatiotemporal orientation is expected to change between views due to the geometry of the situation. In this section, constraint is derived between corresponding stequels subject to rectified and otherwise calibrated binocular viewing. This constraint is derived in two steps. First, the relationship between local spatiotemporal orientations in left and right image spacetime is derived as a 3D scene point $\mathbf{P}$ suffers an arbitrary (infinitesimal) 3D displacement, $\mathbf{V}$, relative to the imaging system. While the relationship between left- and right-based flow has been investigated previously (e.g., [26]), the present derivation sets it in the light of left/right spatiotemporal orientation differences with application to disparity estimation. Second, the left/right flow relationships are generalized to capture the relationship between arbitrary orientations in left and right spacetimes. These results lead directly to the desired relationship between binocular stequels in correspondence.

In the following, bold and regular fonts denote vectors and scalars (resp.), uppercase denotes points relative to the world, lowercase denotes points relative to an image, superscripts l and r denote left and right cameras (resp.), subscripts x, y, z, t specify coordinate components, and vectors in image spacetime taken from time $t = 0$ to $t = 1$ will be distinguished further with tilde. As examples: $\mathbf{P}_t^l = \begin{bmatrix} P_x^l & P_y^l & P_z^l \end{bmatrix}^\top$ is the left camera representation of $\mathbf{P}$ at time t; $\mathbf{p}_t^l = \begin{bmatrix} p_x^l & p_y^l \end{bmatrix}^\top$ is the left image coordinate of $\mathbf{P}_t^l$; $\widetilde{\mathbf{w}} = \begin{bmatrix} w_x & w_y & 1 \end{bmatrix}^\top$ is a vector in image spacetime xyt from $t = 0$ to $t = 1$.

Left-Right Flow Relationship. Consider how a 3D point, $\mathbf{P}$, is observed by the cameras as a function of time, t, while it is displaced along 3D direction, $\mathbf{V}$. The geometry of the situation is shown in Fig. 2. Cameras share a common intrinsic matrix $\mathsf{K} = \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix}$, where other components of the matrix are accounted for by calibration and neglected. At time t, the projections of $\mathbf{P}$ to the left and right views are given by

$$\begin{aligned} \mathbf{P}_t^l &= \mathsf{K}\left((\mathbf{P}_{t=0} - \mathbf{B}) + t\mathbf{V}\right) = \mathbf{P}_{t=0}^l + t\mathsf{K}\mathbf{V} \quad (3) \\ \mathbf{P}_t^r &= \mathsf{K}\left((\mathbf{P}_{t=0} + \mathbf{B}) + t\mathbf{V}\right) = \mathbf{P}_{t=0}^r + t\mathsf{K}\mathbf{V}. \end{aligned}$$

Note that both moving and stationary points are encompassed in this formulation, as $\mathbf{V}$ is arbitrary. The corresponding image coordinates are found in the usual way, *e.g.* for the left view

$$\mathbf{p}^l = \begin{bmatrix} p_x^l \\ p_y^l \end{bmatrix} = \begin{bmatrix} P_x^l / P_z^l \\ P_y^l / P_z^l \end{bmatrix} = \frac{1}{P_z^l} \begin{bmatrix} P_x^l \\ P_y^l \end{bmatrix} = Z^{-1}\mathbf{P}_{2\times 1}^l, \quad (4)$$

where $P_z^l = Z$ is the distance along the Z-axis to the point of regard, $\mathbf{P}$, and $\mathbf{P}_{2\times 1}$ is the upper 2×1 component of $\mathbf{P}$. Analogously for right view, $\mathbf{p}^r = Z^{-1}\mathbf{P}_{2\times 1}^r$.

In the image spacetime coordinate system, xyt, without loss of generality, consider flows $\widetilde{\mathbf{v}}^l$ and $\widetilde{\mathbf{v}}^r$ in the left and right views from temporal instance 0 to 1:

$$\widetilde{\mathbf{v}}^l = \begin{bmatrix} \mathbf{p}_{t=1}^l - \mathbf{p}_{t=0}^l \\ v_t^l \end{bmatrix} = \begin{bmatrix} \mathbf{p}_{t=1}^l - \mathbf{p}_{t=0}^l \\ 1 \end{bmatrix}, \quad (5)$$

where $v_t^l = 1$ by definition, as time has been taken from

$t=0$ to $t=1$. Analogously for the right view

$$\widetilde{\mathbf{v}}^r = \begin{bmatrix} \mathbf{p}^r_{t-1} - \mathbf{p}^r_{t-0} \\ 1 \end{bmatrix}. \quad (6)$$

To relate the left and right spatiotemporal orientations, it is useful to cast the left-camera flow vectors (5) and their right camera counterparts in terms of temporally varying position (3) and (4). Left camera-based flow is given by (5) and substitution from (4) yields

$$\widetilde{\mathbf{v}}^l_{2\times 1} = Z^{-1}_{t=1}\mathbf{P}^l_{2\times 1,t=1} - Z^{-1}_{t=0}\mathbf{P}^l_{2\times 1,t=0}.$$

Further substitution for $\mathbf{P}^l$ according to (3) and letting all subscripts pertain to time (i.e, 0 and 1 denote $t=0$ and $t=1$, resp.) yields

$$\widetilde{\mathbf{v}}^l_{2\times 1} = \frac{Z_0 - Z_1}{Z_0 Z_1}\overline{\mathsf{K}}\mathbf{P}_0 + \frac{1}{Z_1}\overline{\mathsf{K}}\mathbf{V} - \frac{Z_0 - Z_1}{Z_0 Z_1}\overline{\mathsf{K}}\mathbf{B}, \quad (7)$$

where $\overline{\mathsf{K}} = \mathsf{K}_{2\times 3}$ is the top two rows of K. Similarly, for the right camera-based flow

$$\widetilde{\mathbf{v}}^r_{2\times 1} = \frac{Z_0 - Z_1}{Z_0 Z_1}\overline{\mathsf{K}}\mathbf{P}_0 + \frac{1}{Z_1}\overline{\mathsf{K}}\mathbf{V} + \frac{Z_0 - Z_1}{Z_0 Z_1}\overline{\mathsf{K}}\mathbf{B}. \quad (8)$$

Finally, the relationship between the left (7) and right (8) flows is revealed by taking their difference

$$\widetilde{\mathbf{v}}^r - \widetilde{\mathbf{v}}^l = \begin{bmatrix} 2\left(Z_0 - Z_1\right)\overline{\mathsf{K}}\mathbf{B}/\left(Z_0 Z_1\right) \\ 0 \end{bmatrix} = \begin{bmatrix} \Delta \\ 0 \\ 0 \end{bmatrix}, \quad (9)$$

where $\Delta = 2Bf\left(Z_0 - Z_1\right)/\left(Z_0 Z_1\right)$ captures the instantaneous change in disparity.

General Left/Right Orientation Relationship. The relationship (9) was derived only for dominant motion orientation; whereas, stequels capture information from *all* directions $\widetilde{\mathbf{w}}$ in (x, y, t), which now are considered.

Consider directions $\widetilde{\mathbf{w}}^r$ and $\widetilde{\mathbf{w}}^l$ in the left and right views, resp., that are in binocular correspondence, but otherwise arbitrary in (x, y, t). Discounting the effects of right and left flows, $\widetilde{\mathbf{v}}^r$ and $\widetilde{\mathbf{v}}^l$, yields vectors

$$\delta^r = \widetilde{\mathbf{w}}^r - \widetilde{\mathbf{v}}^r = \begin{bmatrix} \delta^r_x & \delta^r_y & 0 \end{bmatrix}^\top, \quad (10)$$

$$\delta^l = \widetilde{\mathbf{w}}^l - \widetilde{\mathbf{v}}^l = \begin{bmatrix} \delta^l_x & \delta^l_y & 0 \end{bmatrix}^\top \quad (11)$$

that capture the purely spatial orientation of corresponding elements (see Fig. 1b,c). For the special case of frontoparallel surfaces $\delta^r = \delta^l$, *i.e.* disregarding motion, oriented texture appears the same across binocular views. For the more general case where surfaces are slanted with respect to the imaging system, the imaged orientation of corresponding elements changes across views, even in the absence of motion. For present matters, this change can be modeled by a linear transformation $\delta^r = \mathsf{A}\delta^l$. Considering that the third element of the δ vectors is always zero by construction, and $\delta^r_y = \delta^l_y$ due to conventional stereo epipolar constraints for rectified setups, this relationship takes the form

$$\delta^r = \mathsf{A}\delta^l, \text{ where } \mathsf{A} = \begin{bmatrix} a_1 & a_2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \quad (12)$$

Substituting (10), (11) into (12) and rearranging yields,

$$\widetilde{\mathbf{w}}^r = \mathsf{A}\widetilde{\mathbf{w}}^l - \mathsf{A}\widetilde{\mathbf{v}}^l + \widetilde{\mathbf{v}}^r. \quad (13)$$

Further substitution of (9) results in

$$\begin{aligned} \widetilde{\mathbf{w}}^r &= \mathsf{A}\widetilde{\mathbf{w}}^l + \left(-\mathsf{A}\widetilde{\mathbf{v}}^l + \widetilde{\mathbf{v}}^l + \begin{bmatrix} \Delta & 0 & 0 \end{bmatrix}^\top\right) \quad (14) \\ &= \begin{bmatrix} a_1 & a_2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\widetilde{\mathbf{w}}^l + \begin{bmatrix} 1-a_1 & -a_2 & \Delta \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}\widetilde{\mathbf{v}}^l \\ &= \begin{bmatrix} a_1 & a_2 & \left((1-a_1)\widetilde{v}^l_x - a_2\widetilde{v}^l_y + \Delta\right) \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\widetilde{\mathbf{w}}^l \end{aligned}$$

Finally, letting $h_1 = a_1 - 1$, $h_2 = a_2$ and $h_3 = \left((1-a_1)\widetilde{v}^l_x - a_2\widetilde{v}^l_y + \Delta\right)$ yields the desired transformation between arbitrary corresponding vectors $\widetilde{\mathbf{w}}^l$ and $\widetilde{\mathbf{w}}^r$

$$\widetilde{\mathbf{w}}^r = \mathsf{H}\widetilde{\mathbf{w}}^l, \text{ where } \mathsf{H} = \begin{bmatrix} 1+h_1 & h_2 & h_3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (15)$$

With (15) in place, it is possible to relate corresponding stequels. By design, (2), stequel Q reveals the amount of intensity variation along all directions in spacetime, and the response ϕ to unit direction $\hat{\mathbf{w}} = \mathbf{w}/\sqrt{\mathbf{w}^T\mathbf{w}}$ is

$$\phi = \hat{\mathbf{w}}^\top \mathsf{Q}\hat{\mathbf{w}}, \quad (16)$$

see, e.g., [10]. Assuming that spatiotemporal correspondences vary in orientation pattern, but not in the intensity per se[1], the responses, ϕ^l, ϕ^r, of corresponding stequels, $\mathsf{Q}^l, \mathsf{Q}^r$, must be the same for related directions, $\hat{\mathbf{w}}^l, \hat{\mathbf{w}}^r$, *i.e.*

$$\hat{\mathbf{w}}^{l\top}\mathsf{Q}^l\hat{\mathbf{w}}^l = \hat{\mathbf{w}}^{r\top}\mathsf{Q}^r\hat{\mathbf{w}}^r.$$

Expanding the normalizations of $\hat{\mathbf{w}}^l$ and $\hat{\mathbf{w}}^r$ and substituting from (15) produces

$$\frac{\widetilde{\mathbf{w}}^{l\top}\mathsf{Q}^l\widetilde{\mathbf{w}}^l}{\widetilde{\mathbf{w}}^{l\top}\widetilde{\mathbf{w}}^l} = \frac{\widetilde{\mathbf{w}}^{l\top}\mathsf{H}^\top\mathsf{Q}^r\mathsf{H}\widetilde{\mathbf{w}}^l}{\widetilde{\mathbf{w}}^{l\top}\mathsf{H}^\top\mathsf{H}\widetilde{\mathbf{w}}^l},$$

while noticing that $\widetilde{\mathbf{w}}^l = \|\widetilde{\mathbf{w}}^l\|\hat{\mathbf{w}}^l$ yields

$$\frac{\hat{\mathbf{w}}^{l\top}\mathsf{Q}^l\hat{\mathbf{w}}^l}{\hat{\mathbf{w}}^{l\top}\hat{\mathbf{w}}^l} = \frac{\hat{\mathbf{w}}^{l\top}\mathsf{H}^\top\mathsf{Q}^r\mathsf{H}\hat{\mathbf{w}}^l}{\hat{\mathbf{w}}^{l\top}\mathsf{H}^\top\mathsf{H}\hat{\mathbf{w}}^l}. \quad (17)$$

Since (17) holds for arbitrary orientations $\hat{\mathbf{w}}^l$ when Q^l and Q^r are stequels in correspondence, it provides the sought for general constraint on binocular stequels. It will be referred to as the *stequel correspondence constraint* and used to derive an approach to stereo matching.

2.3. Stequel match cost

To determine whether two stequels $\mathsf{Q}^l(x, y, t)$ and $\mathsf{Q}^r(x + d, y, t)$ are in correspondence with disparity d, a match cost must be defined. In this section, this cost is derived based on the stequel correspondence constraint, (17), and is taken as the error residual that results from solving for $\mathbf{h} = \begin{bmatrix} h_1 & h_2 & h_3 \end{bmatrix}^\top$ given two candidate stequels.

For a given direction vector $\hat{\mathbf{w}}^l_m$ at some particular orientation m and matching stequels, Q^l and Q^r, the stequel correspondence constraint, (17), yields a quadratic equation

[1] This is a weak form of brightness constancy as any additive and multiplicative intensity offsets between correspondences are compensated for by the bandpass and normalized filters used in stequel construction (2).

in the unknowns of $\mathbf{h}$ of the form

$$f_m(\mathbf{h}) = \left(\hat{\mathbf{w}}_m^{l\top} \mathsf{Q}^l \hat{\mathbf{w}}_m^l\right)\left(\hat{\mathbf{w}}_m^{l\top} \mathsf{H}^\top \mathsf{H} \hat{\mathbf{w}}_m^l\right) - \left(\hat{\mathbf{w}}_m^{l\top} \hat{\mathbf{w}}_m^l\right)\left(\hat{\mathbf{w}}_m^{l\top} \mathsf{H}^\top \mathsf{Q}^r \mathsf{H} \hat{\mathbf{w}}_m^l\right) = 0. \quad (18)$$

Taking a set of M directions, reasonably selected along the same spanning set of directions used to construct Q^l, yields a set of M equations in the three unknowns of $\mathbf{h}$. Thus, $\mathbf{h}$ can be estimated by minimizing a sum of squared errors

$$E_4 = \sum_{m-1}^{M} f_m(\mathbf{h})^2, \quad (19)$$

which is quartic in the entries of $\mathbf{h}$. While such a solution could be sought through analytic or numerical means, it has potential to be expensive to compute and noise sensitive owing to its order. Therefore, it is useful to linearize each error Eqn. (18) through expansion as a Taylor series in $\mathbf{h}$ and retention of terms only through first-order to get

$$g_m(\mathbf{h}) = f_m(\mathbf{0}) + \nabla f_m^\top(\mathbf{0})\mathbf{h}, \quad (20)$$

with $\mathbf{0}$ being the $M \times 1$ zero vector. Using (20), the final function to be minimized with respect to $\mathbf{h}$ becomes

$$E_2 = \sum_{m-1}^{M} \left(f_m(\mathbf{0}) + \nabla f_m^\top(\mathbf{0})\mathbf{h}\right)^2, \quad (21)$$

which is simply quadratic in the elements of $\mathbf{h}$, and thereby can be solved for via standard linear least-squares. More specifically, letting $\mathsf{G} = [\nabla f_1^\top(\mathbf{0}), \nabla f_2^\top(\mathbf{0}), \ldots, \nabla f_M^\top(\mathbf{0})]^\top$ and $\mathbf{c} = -[f_1(\mathbf{0}), f_2(\mathbf{0}), \ldots, f_M(\mathbf{0})]^\top$ yields

$$\mathbf{h} = \left(\mathsf{G}^\top \mathsf{G}\right)^{-1} \mathsf{G}^\top \mathbf{c}; \quad (22)$$

$$E_2 = \|\mathsf{G}\mathbf{h} - \mathbf{c}\|_2^2 = \mathbf{c}^\top \mathbf{c} - \left(\mathsf{G}^\top \mathbf{c}\right)^\top \left(\mathsf{G}^\top \mathsf{G}\right)^{-1} \mathsf{G}^\top \mathbf{c}.$$

For two stequels under consideration for stereo correspondence this residual, E_2, will serve as the local match cost[2].

3. Empirical Evaluation

A software implementation has been developed that inputs a binocular video, computes stequels $\mathsf{Q}^l(x, y, t)$ and $\mathsf{Q}^r(x, y, t)$ for both sequences according to formula (2)[3] and calculates the local match cost, (22), for any given disparity d, i.e., for stequels related as $\mathsf{Q}^l(x, y, t)$ and $\mathsf{Q}^r(x + d, y, t)$. To show the applicability of this approach to disparity estimation, the local match cost, (22), has been embedded in a coarse-to-fine local block-matching algorithm with shiftable windows [22] working over a Gaussian pyramid and also in a global graph-cuts with occlusions matcher [14] operating at the finest scale only; these matchers will be denoted **ST-local** and **ST-global**. Pixel-based disparity estimates are brought to subpixel precision via a Lucas-Kanade type refinement for stequels [1, 23].

To compare with non-stequel matching, versions of the local and global matchers that work simply on single left/right frame pixel comparisons are considered; these matchers will be denoted **noST-local** and **noST-global**, resp. Here, the normalized cross-correlation was used as the data cost term for local and global matching. Finally, to compare to an alternative method for enforcing temporal coherence, optical flow is estimated and used to define a spatiotemporal direction for match cost aggregation that operates over an equivalent number of frames as does the oriented filtering used in stequel construction (1). Here, optical flow is recovered from the stequel representation itself (see [23] and [10] for discussion) to make the comparison fair. The optical flow-based temporal aggregation is used only in conjunction with the local matcher, as incorporation into the global matcher by constructing a spatiotemporal MRF graph [15] is beyond the scope of this paper. The local flow-based aggregation matcher will be denoted **flowAg-local**.

Three data sets are considered. The first is a laboratory sequence (*Lab1*) captured with BumbleBee stereo camera [19] with (framewise) ground truth disparity and discontinuity maps recovered according to a well-known structured light approach [21], see Fig. 3. This scene includes planes slanted in depth with texture oriented along epipolar lines (upper-central part of the scene), various bar-plane arrangement with identical repetitive textures (lower-central part of the scene) and complicated objects with non-trivial 3D boundaries and non-Lambertian materials (e.g., the teddy bear and gargoyle). For this sequence the stereo camera makes a complicated motion that translates along horizontal and depth axes, while part of the scene moves up and down; both camera and scene are on motorized stages.

Visual inspection of the image results (Fig. 3) shows that **noST-local** performs relatively poorly. Planar regions with epipolar aligned texture are generally difficult. Simple temporal aggregation provided by **flowAg-local** is seen to improve on these difficulties; however, performance degrades near 3D boundaries due to unreliable recovery of flow estimates in such areas. **ST-local** does the best of the three local matchers as its *ability to include temporal information allows it to resolve match ambiguities without explicit flow recovery*. As particular improvements of **ST-local** over **noST-local** and **flowAg-local**, consider the lower right and left regions marked with red rectangles in Fig. 3, which highlight the complex outline of the gargoyle wings and the vertical bar in front of plane both having identical textures (camouflage). **ST-local** is quite accurate in these challenging regions, while the other local methods perform relatively poorly. Objects located at different depths in space give rise

[2] Significantly, preliminary experiments showed that match cost based on the linearized error, (21), yielded slightly superior results to considering the original nonlinear error, (19). This can be explained by noting that interest is in a discriminative error measure that reliably penalizes bad matches, and not in the precise error value per se (see [23] for details).

[3] Preliminary experiments considered alternative stequel definition $\mathsf{Q} = \sum \nabla I (\nabla I)^\top$, with $\nabla = [I_x I_y I_t]^\top$ is the spatiotemporal gradient and summation is over local spacetime regions (see [23] for details), but found generally inferior results; so, this approach is considered no further.

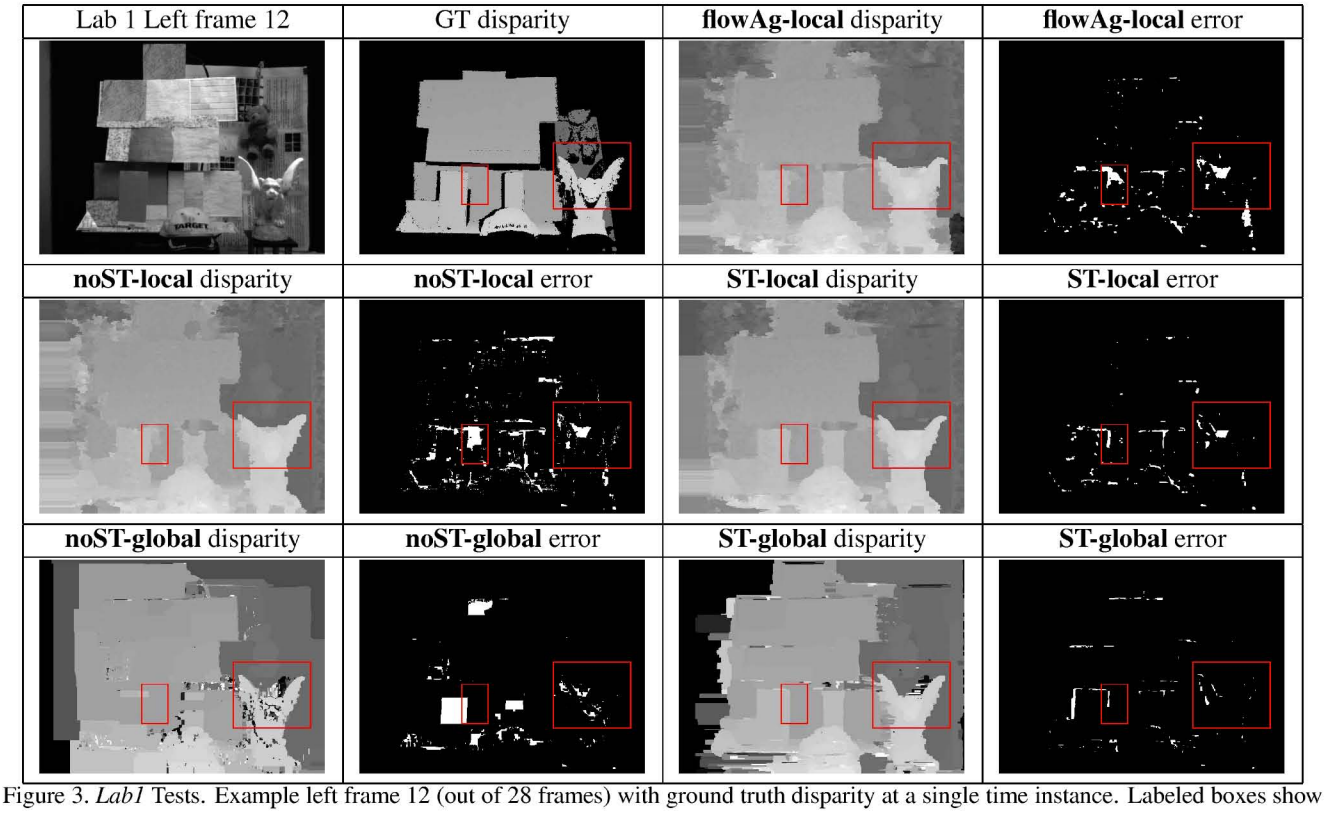


Figure 3. *Lab1* Tests. Example left frame 12 (out of 28 frames) with ground truth disparity at a single time instance. Labeled boxes show recovered disparity maps for compared algorithms and disparity-ground truth absolute differences. A few regions of particular interest in comparing results are highlighted with red rectangles, best seen in color.

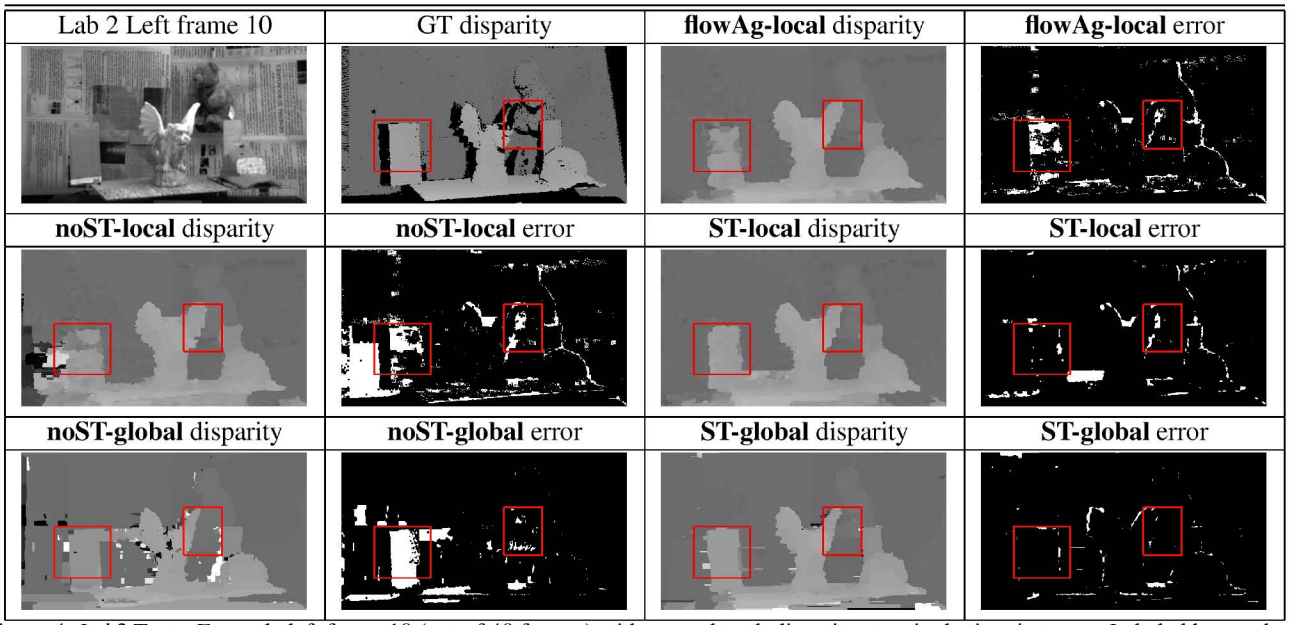


Figure 4. *Lab2* Tests. Example left frame 10 (out of 40 frames) with ground truth disparity at a single time instance. Labeled boxes show recovered disparity maps for compared algorithms and disparity-ground truth absolute differences. A few regions of particular interest in comparing results are highlighted with red rectangles, best seen in color.

to different image motions, even if they undergo the same world motion – and this difference is captured with stequels not allowing for improper matches.

For the global matchers, it is seen even with **noST-global** that it is possible to recover more precisely the complicated 3D boundaries and to achieve good disparity estimates

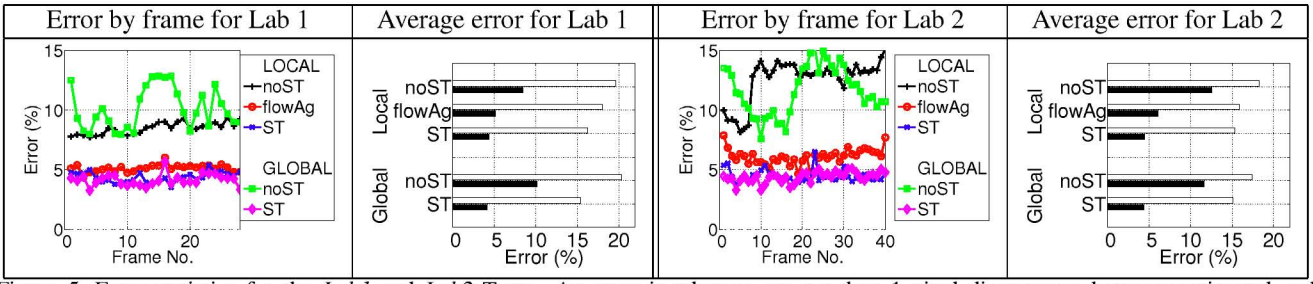


Figure 5. Error statistics for the *Lab1* and *Lab2* Tests. An error is taken as greater than 1 pixel discrepancy between estimated and groundtruth disparity. Bar plots show average error across entire sequences: White bars are for points within 5 pixels of a surface discontinuity; black bars show overall error. Error by frame plots show percentage of points in error overall for each frame separately.

in low texture regions via propagation from better defined boundary matches. However, **noST-global** performs poorly in the regions with epipolar aligned texture and camouflage, as initially incorrect estimates are not subsequently corrected. While increasing the smoothness improves on epipolar-aligned textures, it comes at the expense of camouflage resolution and vice versa. In comparison, **ST-global** is able to recover disparity reliably in these regions, as *once again the stequel representation supports proper resolution of situations that are ambiguous from the purely spatial information*. Another apparent advantage of the **ST-global** is more temporally consistent results – occasional mismatches in **noST-global** can be significantly amplified by propagating into nearby regions.

A second lab sequence, *Lab2*, is constructed in the same controlled environment as *Lab1*, but acquired with significant depth motion and out-of-plane rotation. This particular motion configuration is the most difficult for spatiotemporal stereo, as it results in significantly different left and right spatiotemporal volumes due to slanted surfaces and depth motion. Furthermore, large image motions are present in the individual left and right sequences. Figure 4 presents sample frame results for all five algorithmic instantiations considered above. Here, the conclusions reached from the analysis of *Lab1* are reinforced. With respect to the local methods, **ST-local** provides the most benefit both in weakly textured regions and near 3D boundaries. The performance of **flowAg-local** is hampered by large image motions, which are problematic to recover explicitly in this case; whereas, *direct stequel-based matching is still able to capitalize on temporal information without resolving flow and thereby operates well in the presence of nontrivial motions*. With respect to the global methods, the stequel-based matching **ST-global** significantly outperforms its pixel-based counterpart **noST-global**, especially for weakly-textured highly slanted foreground surfaces.

Error plots for both *Lab1* and *Lab2* quantify the improvements of stequel-based matching in comparison to rivals **noST** and **flowAg** (Fig. 5). Average errors across the sequences show the benefit of stequels near discontinuities and overall for both local and global matchers. Plots of error/frame reinforce the average improvements, but also document improved temporal coherence, as the stequel-based plots vary relatively little across frames, especially in comparison to purely spatial matching provided by **noST**. Incorporation of the temporal dimension also benefits **flowAg**, as its frame-by-frame statistics are relatively stable (albeit overall inferior to stequels); however, the more naturalistic imagery of the next example further emphasizes the superior temporal coherence offered by stequels, even in comparison to **flowAg**.

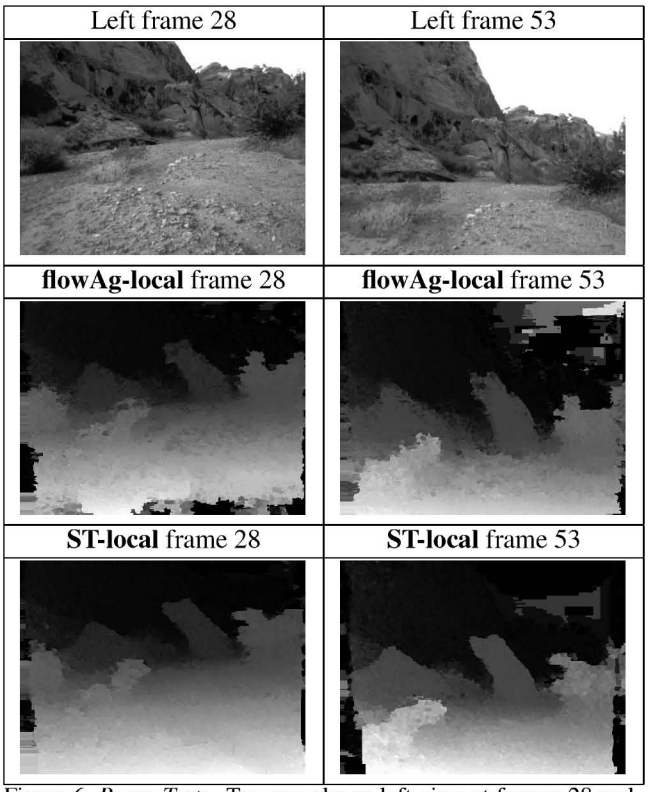


Figure 6. *Rover Tests*. Top row shows left view at frames 28 and 53. Recovered disparity maps at corresponding times are shown below for two algorithms under consideration.

The third data set, *Rover*, is an outdoor sequence acquired from a robot rover traversing rugged terrain, including a receding foreground plane, a central diagonal rock

outcropping, left side cliff, various boulders and bushes. Here the comparison focuses on the improvements to temporal coherence offered by **ST-local** over the rival method for consideration of temporal information, **flowAg-local**. As results of depicted frames show, flow-based aggregation, while providing mostly temporally coherent estimates is inferior at recovery of 3D boundaries (boulders' outlines) and still susceptible to occasional gross errors (*e.g.* on the ground plane) due to errors in the recovered flow. In comparison, stequel-based matching, **ST-local**, does not exhibit such problems, as it uses spatiotemporal information in a more direct and complete way.

4. Discussion

This paper described a novel approach to recovering temporally coherent disparity estimates using stequels as a spatiotemporal matching primitive. Temporal coherence arises naturally, as the primitives and the match cost inherently involve the temporal dimension. Further, matches that are ambiguous when considering only spatial pattern are resolved through the inclusion of temporal information. The stequel matching machinery is simple and involves linear computations only, (22). Thorough experimental evaluation on various datasets shows the benefit of stequel matching as incorporated both in local and global algorithms. Stereo sequences with ground truth have been introduced and are available online for comparison with other algorithms, [23].

A particularly notable benefit of stequel matching is the ability to incorporate temporal information *without* image motion recovery. Optical flow estimation is challenging near 3D boundaries, weakly-textured regions and susceptible to an aperture problem – importantly, this paper demonstrated that stequels are powerful in exactly these situations and provide truly temporally coherent estimates with fewer isolated gross errors. Apparently, stequels allow stereo matching to capitalize on available spatiotemporal structure, even when optical flow recovery is difficult. Further, note that it is non-trivial to model continuity in time with, *e.g.* an MRF prior model as, strictly speaking, temporal graph links have to be defined by flow (as in [15]). Stequels, on the other hand, are directly applicably to standard 2D MRF graphs and their successful performance has been documented in this paper.

In conclusion, a computationally tractable and simple solution to spatiotemporal stereo has been presented, which proved to be very reliable, versatile and robust in practice. Future work will concentrate on exploiting the spatiotemporal profile for explicit non-Lambertian and multi-layer matching, as well as extensions to 3D motion recovery.

Acknowledgements

This work was supported by NSERC and MDA Space Missions. P. Jasiobedski and S. Se provided the *Rover* sequence.

References

[1] S. Baker and I. Matthews. Lucas-Kanade 20 years on: A unifying framework. *IJCV*, 56(3):221–255, 2004.

[2] S. Beauchemin and J. Barron. The computation of optical flow. *ACM Comp. Surv.*, 27:433–467, 1995.

[3] K. Cannons and R. P. Wildes. Spatiotemporal oriented energy features for visual tracking. In *ACCV*, pages 532–543, 2007.

[4] J. Davis, R. Ramamoorthi, and S. Rusinkiewicz. Spacetime stereo: A unifying framework for depth from triangulation. *TPAMI*, 27(2):296–302, 2005.

[5] D. Demirdjian and T. Darrell. Using multiple-hypothesis disparity maps and image velocity for 3D motion estimation. *IJCV*, 47:219–228, 2002.

[6] K. G. Derpanis and J. Gryn. 3-D n-th derivative of Gaussian separable steerable filters. *ICIP*, 3:553–556, 2005.

[7] K. G. Derpanis and R. P. Wildes. Early spatiotemporal grouping with a distributed oriented energy representation. In *CVPR*, 2009.

[8] W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *TPAMI*, 13(9):891–906, 1991.

[9] M. Gong. Enforcing temporal consistency in real-time stereo estimation. In *ECCV*, pages 564–577, 2006.

[10] G. Granlund and H. Knutsson. *Signal Processing for Computer Vision*. Kluwer, 1995.

[11] F. Huguet and F. Devernay. A variational method for scene flow estimation from stereo sequences. In *ICCV*, 2007.

[12] M. Jenkin and J. K. Tsotsos. Applying temporal constraints to the dynamic stereo problem. *CVGIP*, 33:16–32, 1986.

[13] D. G. Jones and J. Malik. A computational framework for determining stereo correspondence from a set of linear spatial filters. In *ECCV*, pages 395–410, 1992.

[14] V. Kolmogorov and R. Zabih. Computing visual correspondence with occlusions using graph cuts. In *ICCV*, pages 508–515, 2001.

[15] E. S. Larsen, P. Mordohai, M. Pollefeys, and H. Fuchs. Temporally consistent reconstruction from multiple video streams. In *ICCV*, pages 1–8, 2007.

[16] S. Malassiotis and M. G. Strintzis. Model-based joint motion and structure estimation from stereo images. *CVIU*, 65(1):79–94, 1997.

[17] R. Mandelbaum, G. Salgian, and H. Sawhney. Correlation-based estimation of ego-motion and structure from motion and stereo. In *ICCV*, pages 544–550, 1999.

[18] J. Neumann and Y. Aloimonos. Spatio-temporal stereo using multiresolution subdivision surfaces. *IJCV*, 47(1-3):181–193, 2002.

[19] Point Grey Research. http://www.ptgrey.com, 2008.

[20] J.-P. Pons, R. Keriven, O. D. Faugeras, and G. Hermosillo. Variational stereovision and 3D scene flow estimation with statistical similarity measures. In *ICCV*, pages 597–602, 2003.

[21] D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. *CVPR*, 1:195–202, 2003.

[22] M. Sizintsev and R. P. Wildes. Efficient stereo with accurate 3-D boundaries. In *BMVC*, pages 237–246, 2006.

[23] M. Sizintsev and R. P. Wildes. Spatiotemporal stereo via spatiotemporal quadric element (stequel) matching. Technical Report CS-2008-04, York University, 2008.

[24] G. P. Stein and A. Shashua. Direct estimation of motion and extended scene structure from a moving stereo rig. *CVPR*, 1:211–218, 1998.

[25] C. Strecha and L. van Gool. Motion-stereo integration for depth estimation. In *ECCV*, pages 170–185, 2002.

[26] A. M. Waxman and J. H. Duncan. Binocular image flows: Steps toward stereo-motion fusion. *TPAMI*, 8(6):715–729, 1986.

[27] O. Williams, M. Isard, and J. MacCormick. Estimating disparity and occlusions in stereo video sequences. *CVPR*, pages 250–257, 2005.

[28] L. Zhang, B. Curless, and S. M. Seitz. Spacetime stereo: Shape recovery for dynamic scenes. In *CVPR*, pages 367–374, 2003.

[29] Z. Zhang and O. D. Faugeras. Three-dimensional motion computation and object segmentation in a long sequence of stereo frames. *IJCV*, 7(3):211–241, 1992.