Spatiotemporal Stereo and Scene Flow via Stequel Matching

Mikhail Sizintsev, *Student Member, IEEE,* and Richard P. Wildes, *Member, IEEE,*

Abstract—This paper is concerned with the recovery of temporally coherent estimates of 3D structure and motion of a dynamic scene from a sequence of binocular stereo images. A novel approach is presented based on matching of spatiotemporal quadric elements (**stequels**) between views, as this primitive encapsulates both spatial and temporal image structure for 3D estimation. Match constraints are developed for bringing stequels into correspondence across binocular views. With correspondence established, temporally coherent disparity estimates are obtained without explicit motion recovery. Further, the matched stequels also will be shown to support direct recovery of scene flow estimates. Extensive algorithmic evaluation with ground truth data incorporated in both local and global correspondence paradigms shows the considerable benefit of using stequels as a matching primitive and its advantages in comparison to alternative methods of enforcing temporal coherence in disparity estimation. Additional experiments document the usefulness of stequel matching for 3D scene flow estimation.

Index Terms—Stereo, motion, spacetime, spatiotemporal, scene flow, quadric element, stequel.

1 Introduction

1.1 Motivation

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 considerably less attention. Ultimately, recovery of 3D scene structure must respect dynamic information to ensure that estimates are temporally consistent. Similarly, 3D motion estimates must be consistent with scene structure. Moreover, in situations where instantaneous binocular 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 as well as 3D scene flow 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 referred to as the orientation tensor and covariance matrix, see, e.g., [20], [3]). By representing orientation structure uniformly across image space and time, both instantaneously defined (e.g., spatial texture) and dynamically defined (e.g., motion) information can be brought to bear on stereo correspondence in an integrated fashion. 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. At the same time, 3D scene flow vectors can be computed directly from the matched spatiotemporal primitives. Further, this representation allows spatial and temporal image structure to resolve otherwise ambiguous matches in a fashion consistent with both sources of information. Significantly, applicability of this representation to stereo correspondence is quite general and will be demonstrated in both local and global matchers.

• *M. Sizintsev and R.P. Wildes are with the Department of Computer Science and Engineering and Centre for Vision Research, York University, Toronto, Ontario, Canada.*
E-mail: see http://www.cse.yorku.ca/vision

1.2 Previous work

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 [25]. A rather different early approach exploited constraints on the temporal derivative of disparity [53], based on an earlier psychophysically motivated analysis [39]. Other work matched binocular features to recover 3D estimates for temporal tracking [54], [59]. More recent research that relies on loose coupling of stereo and motion has placed greater emphasis on recovery of dense estimates. One representative approach begins by recovery of binocular matches, followed by joint recovery of consistent left and right optical flows for final combination into 3D flow [58]. The general idea of combining disparity with left-right consistent optical flows also has been extended to consider space carving [30] coupled with simultaneous estimation of optical flow from larger collections of cameras [17]. A rather different approach emphasized the recovery of 3D motion using optical flow in conjunction with multiple

hypothesis binocular disparity maps [8]. The proposed research differs from all such work in being focused on a more integrated approach to spatiotemporal processing.

More recent stereo research has seen increased interest in scene recovery from multi-camera (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 [33]. 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 [36]. 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 [38]. Again, the proposed approach differs from these lines of research in its 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 processing. Some of this work has concentrated on static scenes with variable lighting [7]. Others have focused on defining appropriate temporal integration windows, e.g., as part of the correspondence process [57] or simply reinforce disparity estimates from the previous frame using optical flow [19]. Further, combined stereo and motion estimation has been formulated in terms of both PDEs [50], [23] and MRFs [51], [32], [55], [24], [31]. Still other work has used direct methods for integrated recovery of structure and egomotion [21], [48], [34]. Yet other approaches have formulated matters as a 6D estimation problem (3D position and 3D velocity associated with it) by fusing stereo and optical flow recovery into a single estimation [17] as well as through extension of space-carving to spacetime [52]. Finally, approaches have considered infinitesimal motion and stereo disparity as encapsulated in a single brightness constancy equation [40], [41]. 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 employed in previous methods, as they typically worked with 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 has documented enhancement [15], [20], optical flow recovery [1], [22], tracking [5], grouping and segmentation [13], dynamic texture analysis [10] and action recognition [6], [14], [27], [9] on the basis of filters tuned for local spatiotemporal orientation. More specifically, previous research has considered the use of the spatiotemporal quadric to capture orientation in image spacetime, with application to motion estimation, restoration, enhancement [20], [3] and flow comparison [43]. However, it appears none has exploited spatiotemporal orientation, in general, or the spatiotemporal quadric, specifically, for stereo disparity estimation or for scene flow estimation. Previous stereo work has defined binocular correspondence based on a bank of spatial filters [26]. 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 spatial and temporal domains, thereby basing matching on a fundamentally richer representation.

1.3 Contributions

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 spacetime 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) A method for recovery of 3D scene flow is presented based on left-right spatiotemporal quadric correspondences. (iv) Extensive empirical evaluation of resulting disparity and scene flow estimation algorithms is presented. Testing encompasses quantitative evaluation on laboratory acquired binocular video with ground truth and qualitative evaluation on more naturalistic imagery. The laboratory imagery and associated ground truth are available for download [44]. A preliminary version of this research has appeared previously [46].

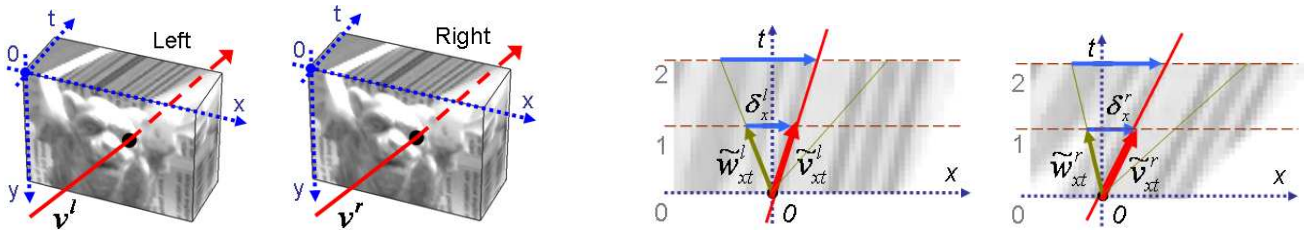
2 Technical Approach

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 pixel intensities, consideration of local spatiotemporal orientation provides access to a richer representation. Local orientation has visual significance as 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, subsequent matching will be inherently constrained to observe temporal coherence. Further, through combination of both temporal and spatial structure in the descriptor, match ambiguities that might exist through consideration of only one data source have potential to be resolved.

2.1.1 3D steerable filters

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



(a) (b) (c)

Fig. 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. $\widetilde{v}^l_{xt}$ and $\widetilde{v}^r_{xt}$ 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.

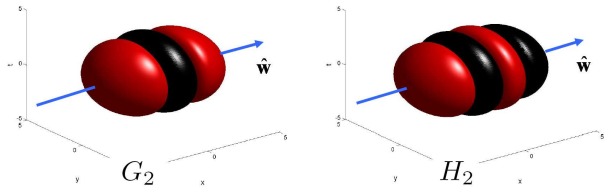


Fig. 2. Surfaces plots of 3D steerable filter pair G_2 and H_2 oriented along the x-axis in the spacetime volume, i.e. $\hat{\mathbf{w}} = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^\top$. Red and black colours represent positive and negative contributions, respectively.

Hilbert transforms, H_2 [18], 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 [18].

Figure 2 visualizes G_2 along a particular direction and its $90°$-phase counterpart H_2 filter. The composed response as in (1) will be phase-invariant and specific to the chosen direction. Furthermore, even very oblique orientations which correspond to large motions can be sampled with G_2-H_2 pairs, and therefore exploited in binocular image matching.

Filtering is applied separately to the left and right image sequences. Here, filters are oriented along normals to icosahedron faces with antipodal directions identified (10 directions in total), as this uniformly samples the sphere and spans 3D orientation for the employed filters. Mathematically, these directions are defined by vectors

$$[\pm 1, \pm 1, \pm 1], \left[0, \frac{\pm 1}{\phi}, \pm\phi\right], \left[\frac{\pm 1}{\phi}, \pm\phi, 0\right], \left[\pm\phi, 0, \frac{\pm 1}{\phi}\right],$$

where $\phi = \frac{\sqrt{5}+1}{2}$, subject to normalization of each vector to unit length. After filtering, every point in spacetime has an associated set of values that indicate how strongly oriented the local structure is along each considered direction.

2.1.2 Constructing the match primitive

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 [20]. 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,

$$\mathsf{Q} = \sum_i \hat{E}_i \left(\frac{5}{4} \hat{\mathbf{w}}_i \hat{\mathbf{w}}_i^\top - \frac{1}{4} \mathsf{I}_3 \right), \quad (2)$$

where I_3 is an identity matrix, the summation is across the set of filter orientations, $\hat{\mathbf{w}}_i$, and $\hat{E}_i$ is the corresponding local energy response (1), but now normalized such that $\sum_i \hat{E}_i(\mathbf{x}) = 1$ (see Sec. 2.2 for normalization rationale). In constructing Q, the dyadic product, $\hat{\mathbf{w}}_i \hat{\mathbf{w}}_i^\top$, establishes the local frame implied by orientation $\hat{\mathbf{w}}_i$ weighted by its corresponding response, $\hat{E}_i$. Subtraction of the identity component is necessary to remove the bias that otherwise contaminates the local estimate of the quadric [20].

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 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 nothing more than 3D separable convolution and pointwise nonlinearities and is thereby amenable to compact, efficient implementation [12] including real-time realizations on parallel hardware, e.g., GPUs [56]. Construction of Q from the filter responses requires only matrix summation, as specified in (2). Nevertheless, depending on the observed efficacy of this particular filtering approach, alternatives may also be considered: As examples, Gabor and lognormal filters may be considered.

Finally, it is worth noting that the stequel, Q, can be constructed in an alternative fashion using first-order derivatives, i.e., as instantiated via the Grammian, $\sum \nabla I (\nabla I)^\top$

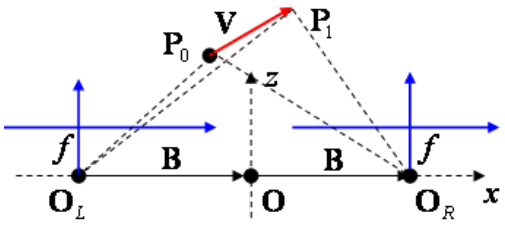


Fig. 3. Stereo Geometry. A reference Euclidean coordinate system is centred at the midpoint of the stereo baseline, $\mathbf{O}$. Cameras are rectified with a half-baseline vector $\mathbf{B} = [b, 0, 0]^\top$ and focal lengths f. Left and right optical centres are at $O^l = -\mathbf{B}$ and $O^r = \mathbf{B}$, resp. Point $\mathbf{P}$ undergoes an arbitrary displacement $\mathbf{V}$ from instance 0 to 1.

with $\nabla I = (I_x, I_y, I_t)^\top$ the spatiotemporal gradient and summation taken over local spatiotemporal regions, e.g., as used in [43] in a different context. While this definition entails lower-order derivatives than (2), preliminary experiments indicated that it yields generally inferior quantitative results in the present context and will not be considered further in experiments (see [45] for details). The experimental advantage of the stequel vs. Grammian can be explained as follows. First, the stequel is constructed from $G2$ and $H2$ filter responses which are more finely tuned to orientation that $G1$ used in Grammian construction. Second, stequels constructed from quadrature pair $G2$ and $H2$ provide a more localized measure of orientation structure as they are pointwise phase invariant; whereas, construction via the Grammian relies on neighborhood aggregation to annihilate phase.

2.2 Spatiotemporal epipolar correspondence 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. Here, displacement can come about through movement of the point, the imaging system or a combination thereof. Further, since the analysis is point-based, no scene rigidity is assumed.

While the relationship between left- and right-based flow has been investigated previously (e.g., [53]), the present derivation sets it in the light of left/right spatiotemporal orientation differences with application to disparity estimation; whereas, previous work assumed disparity estimation and was focused on subsequent 3D inferences. Further, 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 $t = 1$.

2.2.1 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. 3. 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

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

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

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

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

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

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

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

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

2.2.2 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 fronto-parallel 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

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

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 $\widetilde{\mathbf{w}}^r$

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

It is interesting to outline a special case associated with (15). The situation of $h_1 = h_2 = 0$ means that $\delta^r = \delta^l$ in (12), which essentially implies the fronto-parallel assumption, that is still widely used in contemporary stereo matching. This case is quite important from a practical point of view, because it yields reasonable results and can be faster as well as more numerically stable in estimation owing to its simpler form.

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., [20]. 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, $\hat{\mathbf{w}}^l, \hat{\mathbf{w}}^r$:

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

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

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

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

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

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

which is quartic in the entries of $\mathbf{h}$. While such a solution could be sought through analytic or numerical means, it is 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

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

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

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

which is simply quadratic in the elements of $\mathbf{h}$, and thereby can be solved for via standard linear least-squares [49]. More specifically, letting

$$\begin{aligned} \mathsf{G} &= [\nabla f_1^\top(\mathbf{0}), \nabla f_2^\top(\mathbf{0}), \ldots, \nabla f_M^\top(\mathbf{0})]^\top \\ \mathbf{c} &= -[f_1(\mathbf{0}), f_2(\mathbf{0}), \ldots, f_M(\mathbf{0})]^\top \end{aligned}$$

yields

$$\begin{aligned} \mathbf{h} &= \left(\mathsf{G}^\top\mathsf{G}\right)^{-1}\mathsf{G}^\top\mathbf{c}; \qquad (22) \\ E_2 &= \|\mathsf{G}\mathbf{h} - \mathbf{c}\|_2^2 = \mathbf{c}^\top\mathbf{c} - \left(\mathsf{G}^\top\mathbf{c}\right)^\top\left(\mathsf{G}^\top\mathsf{G}\right)^{-1}\mathsf{G}^\top\mathbf{c}. \end{aligned}$$

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

Significantly, preliminary experiments showed that match cost based on the linearized error, (21), yielded only slightly inferior results in comparison to the original nonlinear error, (19), which lends further support to pursuing the advocated (linearized) approach. The nonlinear estimation was performed via Gauss-Newton optimization using the solution of (21) as a starting point. In the context of the current experiment setting described in Sec. 3, the use of the original nonlinear error, (19), vs. linearized cost, (21), results only in 3% relative reduction of overall disparity estimation error; however, it increases the computation time by an order of magnitude. The relatively strong performance of the linearized solution can be explained by the fact that interest is in a *discriminative* error measure that reliably penalizes bad matches, and not in the precise error value per se. Finally, since matching must be done for every point, it must be sufficiently simple to be practical: solution (22) requires the inverse of a 3x3 matrix, which can be coded in closed form; indeed, the whole matching procedure is comparable to normalized cross correlation in terms of computational complexity and runtime.

2.4 Scene Flow Estimation

The results derived so far show how it is possible to use spatiotemporal information to constrain disparity estimation without explicit recovery of motion via the stequel correspondence constraint, (17), and the related match cost, (21). In this section it is shown how to use matched stequels to estimate 3D scene motion directly without using left- and right-based optical flow as an intermediary. In contrast, the left and right stequels could be used independently to recover optical flow for both the left and right image streams, i.e., given a region contains adequate structure, by projecting the eigenvector of the stequels smallest eigenvalue, which captures locally dominant spatiotemporal orientation, onto the image plane [20]. For example, if $\hat{\mathbf{e}}_3 = \begin{bmatrix} e_x & e_y & e_t \end{bmatrix}^\top$, corresponds to the smallest eigenvalue of Q^l, then the left optical flow is given as

$$\begin{bmatrix} e_x/e_t & e_y/e_t \end{bmatrix}^\top. \tag{23}$$

Subsequently, left and right flows could be combined with disparity to yield 3D flow (e.g., analogous to various methods reviewed in Sec. 1.2); however, the method presented below provides more direct access to 3D flow.

Provided camera calibration, spatiotemporal stereo disparity estimates afford the recovery of 4D scene spacetime, $\mathbf{P} = \begin{bmatrix} P_x & P_y & P_z & t \end{bmatrix}^\top$, and scene flow arises as the displacement of a point's spatial position, (P_x, P_y, P_z), with time, t. Since stequels are defined in image spacetime, it is convenient for present purposes to consider the correlate disparity spacetime, $\mathbf{p} = \begin{bmatrix} p_x & p_y & d & t \end{bmatrix}^\top$ with disparity, d, relating matched stequels. Notice that $\mathbf{P}$ and $\mathbf{p}$ are in one-to-one correspondence given calibration. In particular, let disparity spacetime be taken relative to the left camera[2], i.e., consider $\begin{bmatrix} p_x^l & p_y^l & d & t \end{bmatrix}^\top$. Each point projects into the left and right spatiotemporal volumes as

$$\begin{aligned} \mathbf{p}^l &= \Pi^l \begin{bmatrix} p_x^l & p_y^l & d & t \end{bmatrix}^\top = \begin{bmatrix} p_x^l & p_y^l & t \end{bmatrix}^\top \\ \mathbf{p}^r &= \Pi^r \begin{bmatrix} p_x^l & p_y^l & d & t \end{bmatrix}^\top = \begin{bmatrix} p_x^l + d & p_y^l & t \end{bmatrix}^\top, \end{aligned}$$

respectively, with

$$\Pi^l = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \text{ and } \Pi^r = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{24}$$

Analogously, appeal to the left, $\tilde{\mathbf{v}}^l$, and right, $\tilde{\mathbf{v}}^r$, flow definitions, (7) and (8), respectively, allows for specification of their relationships to the disparity spacetime flow, $\tilde{\mathbf{v}} = \begin{bmatrix} v_x & v_y & v_d & 1 \end{bmatrix}^\top$, to be given as

$$\begin{aligned} \tilde{\mathbf{v}}^l &= \begin{bmatrix} v_x & v_y & 1 \end{bmatrix}^\top = \Pi^l\tilde{\mathbf{v}}, \qquad (25) \\ \tilde{\mathbf{v}}^r &= \begin{bmatrix} v_x + v_d & v_y & 1 \end{bmatrix}^\top = \Pi^r\tilde{\mathbf{v}}. \qquad (26) \end{aligned}$$

Here, reference still is to the left camera system; however, superscripts on the components of $\mathbf{v}$ are suppressed for compactness of notation.

To relate the left and right flows to their stequels, recall (as noted above) that the locally dominant spatiotemporal

2. Alternatively, right-based or cyclopean-based camera systems could be considered in a similar fashion.

orientation is captured as the eigenvector of the smallest eigenvalue of the local stequel. Thus,

$$\tilde{\mathbf{v}}^l = \arg\min_{||\mathbf{q}||^2=1} ||\mathsf{Q}^l\mathbf{q}||^2 \text{ and } \tilde{\mathbf{v}}^r = \arg\min_{||\mathbf{q}||^2=1} ||\mathsf{Q}^l\mathbf{q}||^2, \quad (27)$$

where $||\cdot||$ denotes vector length and $\mathbf{q}$ is a dummy variable considered independently in the two expressions. Substitution from the equations that relate $\tilde{\mathbf{v}}$ to $\tilde{\mathbf{v}}^l$ and $\tilde{\mathbf{v}}^r$, (25) and (26), resp., allows the minimizations, (27), to be combined as

$$\tilde{\mathbf{v}} = \arg\min_{||\mathbf{q}||^2=1} \left[||\mathsf{Q}^l\Pi^l\mathbf{q}||^2 + ||\mathsf{Q}^r\Pi^r\mathbf{q}||^2\right].$$

Further expansion of the square and explicit normalization allows the previous equation to be rewritten as

$$\tilde{\mathbf{v}} = \arg\min_{\mathbf{q}} \frac{\mathbf{q}^\top \left(\Pi^{l\top}\mathsf{Q}^{l\top}\mathsf{Q}^l\Pi^l + \Pi^{r\top}\mathsf{Q}^{r\top}\mathsf{Q}^r\Pi^r\right)\mathbf{q}}{\mathbf{q}^\top\mathbf{q}},$$

or more compactly as

$$\tilde{\mathbf{v}} = \arg\min_{\mathbf{q}} \frac{\mathbf{q}^\top\mathsf{Q}_\mathcal{M}\mathbf{q}}{\mathbf{q}^\top\mathbf{q}}, \quad (28)$$

where $\mathsf{Q}_\mathcal{M} = \left(\Pi^{l\top}\mathsf{Q}^{l\top}\mathsf{Q}^l\Pi^l + \Pi^{r\top}\mathsf{Q}^{r\top}\mathsf{Q}^r\Pi^r\right)$ is a 4×4 quadric.

The last statement is a standard Rayleigh quotient [29], which is solved by finding the eigenvector, $\hat{\mathbf{e}}_4 = [e_x\ e_y\ e_d\ e_t]^\top = \tilde{\mathbf{v}}$, associated with the smallest eigenvalue of $\mathsf{Q}_\mathcal{M}$. Finally, the recovery of 3D flow vector vector, $\mathbf{v}_3$, from the 4D eigenvector, $\hat{\mathbf{e}}_4$, proceeds by projecting into $(x,y,d)^\top$-space:

$$\mathbf{v}_3 = \left[\ e_x/e_t \quad e_y/e_t \quad e_d/e_t\ \right]^\top, \quad (29)$$

in a fashion exactly analogous to projecting the 3D eigenvector, $\hat{\mathbf{e}}_3$, to optical flow (23).

While (29) describes an unambiguous way of recovering the 3D motion from two matched stequels, it is important to note that the underlying spatiotemporal structure can limit what can be done in practice. For example, as an extreme case, if a region is completely lacking in structure (i.e., uniform intensity/textureless), then no motion recovery is supported. Happily, the nature of $\mathsf{Q}_\mathcal{M}$'s eignevalues, $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \lambda_4$, allow the situation to be diagnosed (c.f., [20], [35]). The case of a well-defined (flow) vector corresponds to the condition $\lambda_3 >> \lambda_4 \approx 0$, i.e., when the nullspace of $\mathsf{Q}_\mathcal{M}$ is one-dimensional. A measure ζ that reflects this is

$$\zeta = 3\left(\lambda_3 - \lambda_4\right), \quad (30)$$

where $\zeta \in [0,1]$, given that $max(\lambda_3) = 1/3$ is achieved when $\lambda_1 = \lambda_2 = \lambda_3$, for trace-normalized $\mathsf{Q}_\mathcal{M}$. Similarly, a higher dimensional nullspace indicates how underconstrained the flow is, i.e., instances of the aperture problem where only the normal flow might be recovered.

3 Empirical Evaluation

3.1 Algorithmic instantiations

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) 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 [47] working over a Gaussian pyramid and also in a global graph-cuts with occlusions matcher [28] operating at the finest scale only; these matchers will be denoted **ST-local** and **ST-global**. Stequels were constructed from the steerable filter responses, with filters as reported elsewhere [12]. The spatiotemporal support employed for stequel computation was $x \times y \times t = 5 \times 5 \times 5$. Pixel-based disparity estimates are brought to subpixel precision via Lucas-Kanade type refinement for stequels [2], [45]. Further, given matched stequels, Q^l and Q^r, an estimate of 3D scene flow is obtained via algorithmic instantiation of the motion recovery equations (28)-(29). As a representative run-time: Subpixel disparity estimates were recovered at 2 frames/second on 640×480 video for an unoptimized C++ implementation executing on a 3 GHz processor; speed scales linearly with pixel dimensions.

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, zero-mean normalized cross-correlation with 5×5 spatial aggregation was used for local matching. For global pixel matching, the data cost was computed on band-passed images in order to be more robust to radiometric differences between the images (level 0 in a Laplacian pyramid) in stereo pairs. 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, (23), 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 [31] is beyond the scope of this paper. The local flow-based aggregation matcher will be denoted **flowAg-local**. Finally, comparison is made to an alternative spacetime matching technique that uses intensities directly [57]. In essence, this last approach extends block-based matching to the temporal domain by performing aggregation over local spacetime 3D windows with optimization for spatial slant and depth motion. To allow fair comparison, this paradigm was embedded in the aforementioned local [47] and global [28] methods, which will be denoted as **Zhang-local** and **Zhang-global**, resp.

In general, the comparison of local methods is important,

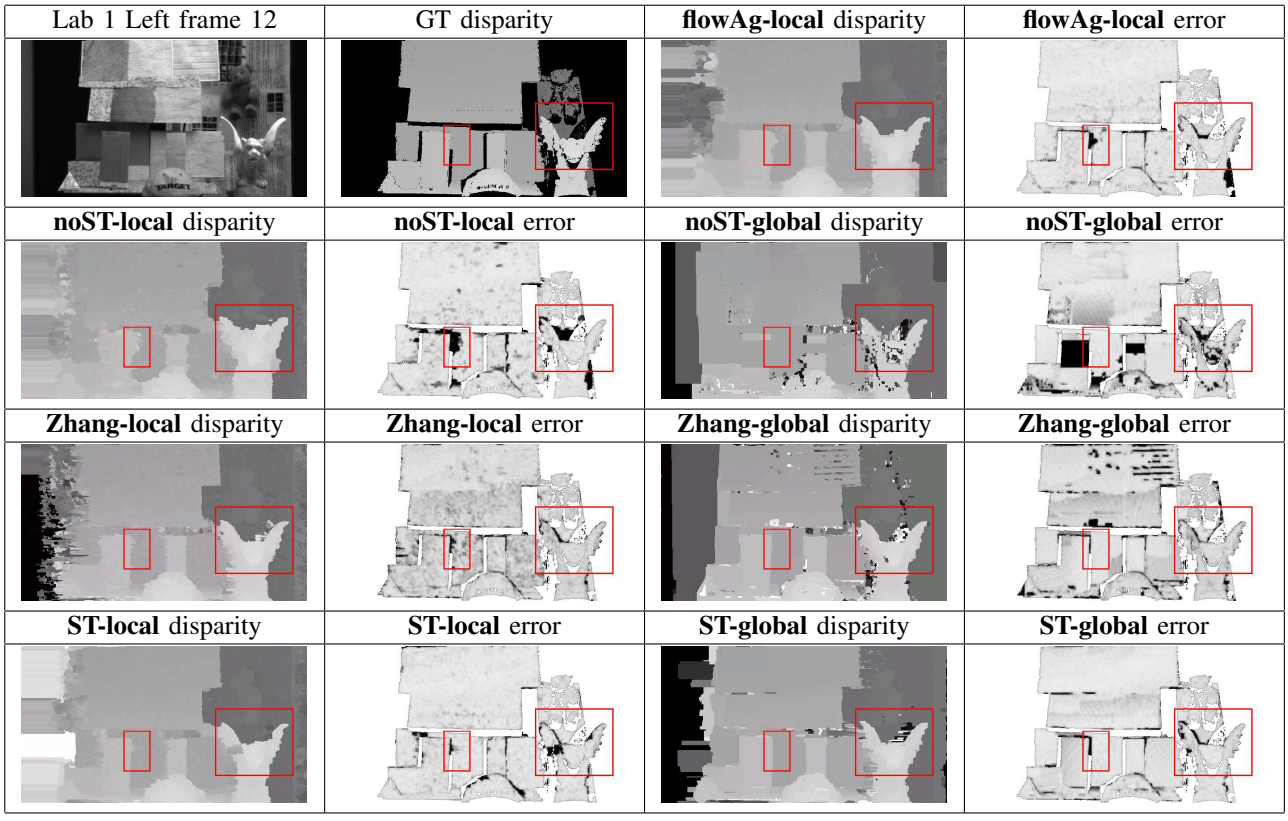


Fig. 4. *Lab1* Tests. Example left and right frame 12 ($x \times y \times t = 640 \times 480 \times 28$) with ground truth disparity (top row). Labeled boxes (beneath) show recovered disparity maps for compared algorithms and disparity-ground truth absolute differences. Error values are truncated at 7 pixels to emphasize small values; brighter intensity correspond to smaller error. A few regions of particular interest in comparing results are highlighted with red rectangles, best seen in color.

as their results depend the most on the quality of matching primitives and, thus, would allow access to the performance of stequel matching in the absence of other cues. The comparison of global methods is crucial, as they generally provide superior results and stequels must be able to show additional benefits in order to be useful in practice.

3.2 *Lab* sequences

Two laboratory data sets are considered. The first is a sequence (*Lab1*) captured with BumbleBee stereo camera [37] with (framewise) ground truth disparity and discontinuity maps recovered according to a well-known structured light approach [42], see Fig. 4. 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. 4) 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. Similarly, **Zhang-local** helps to disambiguate matches in the camouflage region and slightly improves estimates at the epipolar-aligned textured regions. However, these positive aspects of **Zhang-local** are offset by very pronounced errors around image boundaries, which makes its results very temporally-inconsistent and quantitatively quite poor. **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. 4, 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 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 in

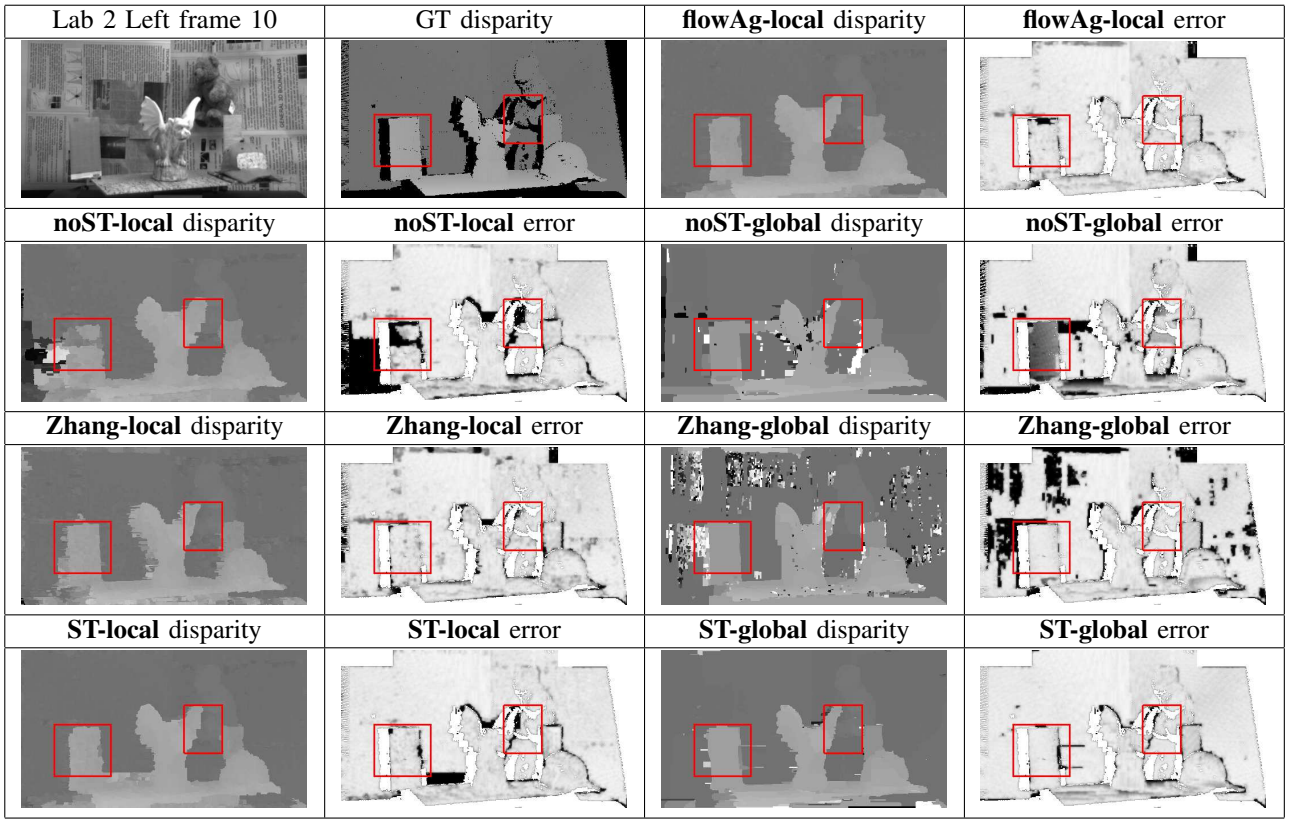


Fig. 5. *Lab2* Tests. Example left and right frame 10 ($x \times y \times t = 640 \times 480 \times 40$) with ground truth disparity (top row). Labeled boxes (beneath) show recovered disparity maps for compared algorithms and disparity-ground truth absolute differences. Error values are truncated at 7 pixels to emphasize small values; brighter intensity correspond to smaller error. A few regions of particular interest in comparing results are highlighted with red rectangles, best seen in color.

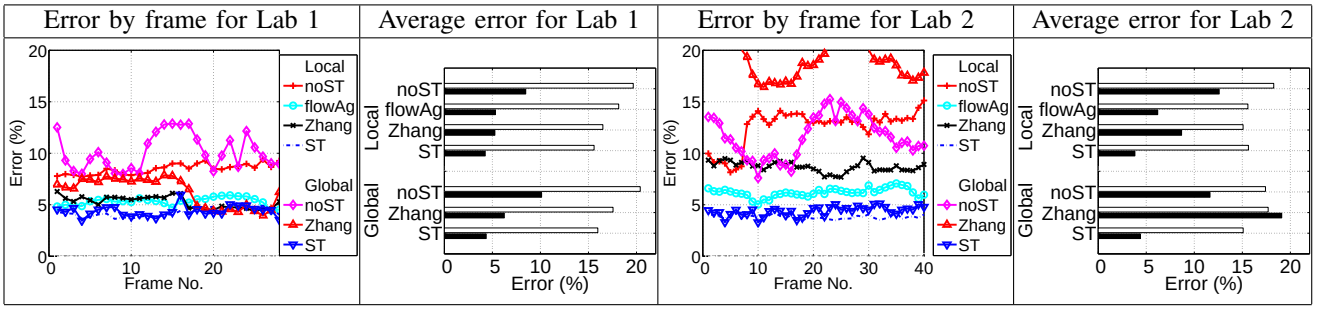


Fig. 6. Error statistics for the *Lab1* and *Lab2* Tests. An error is taken as greater than 1 pixel discrepancy between estimated and groundtruth disparity. Bar plots show average error across entire sequences: White bars are for points within 5 pixels of a surface discontinuity; black bars show overall error. Error by frame plots show percentage of points in error overall for each frame separately.

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 *the stequel representation supports proper resolution of situations that are ambiguous from the purely spatial information*. Note that while **Zhang-global** can deal with the camouflage effect and outperform the **noST-global** overall, it still fails to correctly estimate disparity for simple epipolar-aligned textures, as, once again, incorrect matches are propagated without correction. 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

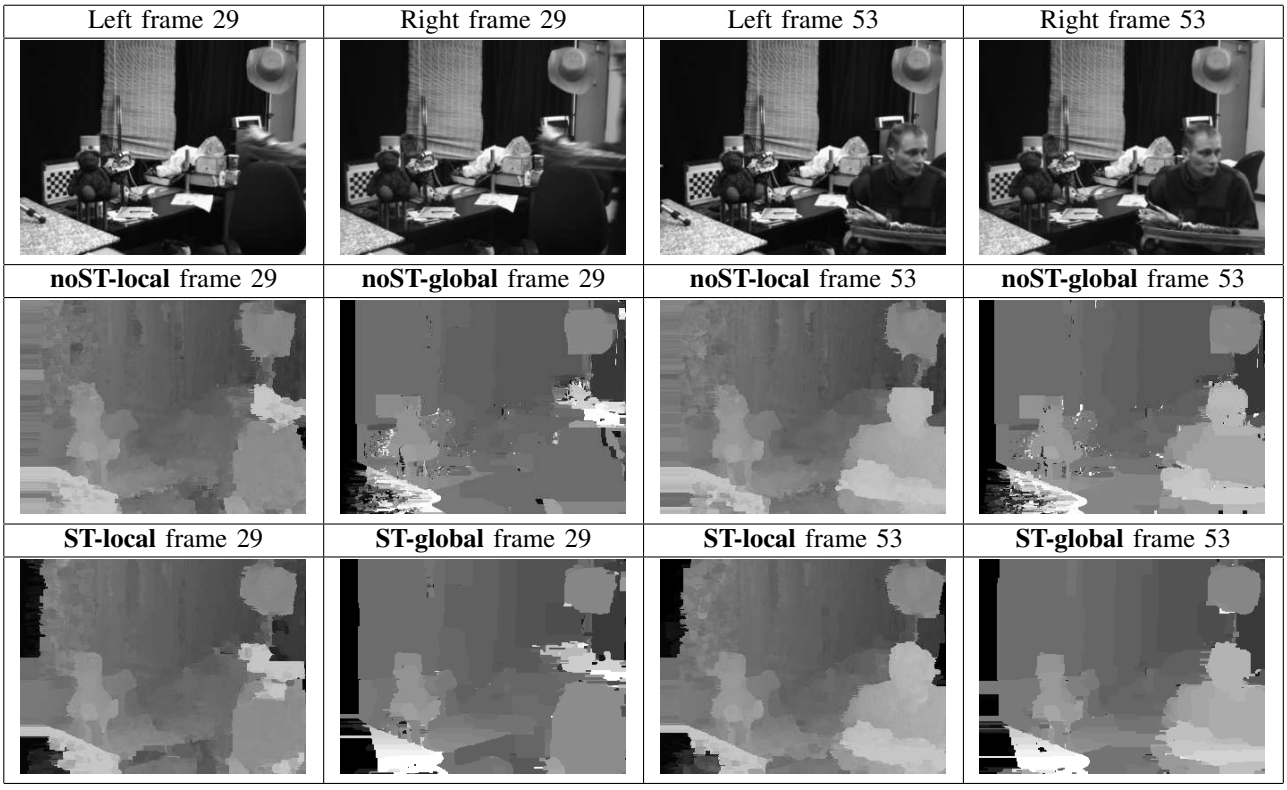


Fig. 7. *Office* Tests. Right column shows left and right images for frames 29 and 53 $(x \times y \times t = 320 \times 240 \times 128)$. Remaining boxes are labeled with recovered disparity by algorithm and frame. See accompanying videos at [44].

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 5 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. In this light it is important to note the particularly poor performance of **Zhang** methods on the fine-textured background. An explanation of this phenomenon is the presence of the zooming effect associated with in-depth motion, which is not effectively captured by the simple temporal window shifts adopted in [57]. In contrast, stequels are constructed as the pointwise measurement of the first-order intensity structure and explicit temporal aggregation is not performed during their matching; hence, no such problem arises.

Error plots for both *Lab1* and *Lab2* quantify the improvements of stequel-based matching in comparison to rivals **noST**, **flowAg** and **Zhang** (Fig. 6). 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 following examples further emphasizes the superior temporal coherence offered by stequels, even in comparison to **flowAg**.

It is worth noting that there are slight differences between results presented for **ST** and **flowAg** in this section and those presented in a previous report on stequel-based disparity estimation [45]; although, these differences are never more than approximately 1% in average error, e.g., as reported in Fig. 6. These differences arise because a different definition for the stequel is used. The earlier report approximated the stequel as $E\sum \mathbf{w}\mathbf{w}^\top$; whereas, now the exact formulation, (2), is employed. This change was motivated by the need for more precise representation to support 3D scene flow estimation, which was not considered previously.

3.3 *Office* sequence

The third data set, *Office*, depicts a more naturalistic (albeit without ground truth) cluttered indoor office scene where the camera pans while a person enters and subsequently moves about in a nonrigid fashion, see Fig. 7. Here, the

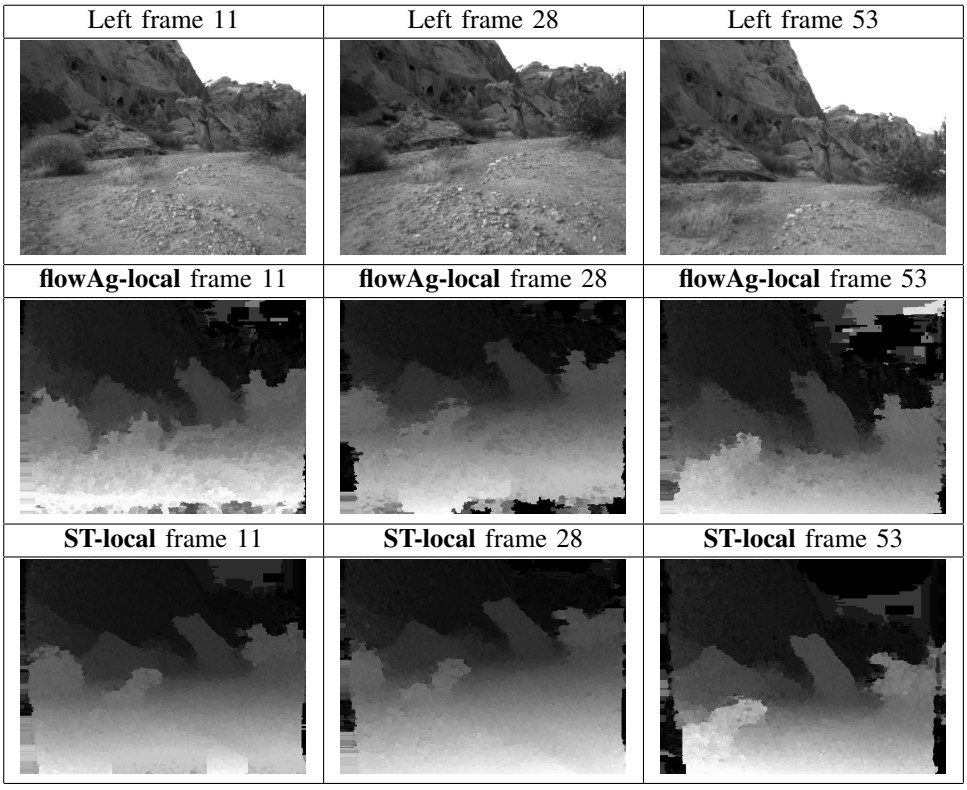


Fig. 8. *Rover Tests*. Top row shows left view at frames 88, 105 and 130 $(x \times y \times t = 640 \times 480 \times 78)$. Recovered disparity maps at corresponding times are shown below for two algorithms. See accompanying videos at [44].

superior ability of stequel matching to produce temporally coherent disparity maps is illustrated via a comparison of stequel and single frame matching; it is seen that both **ST-local** and **ST-global** best their non-stequel-based counterparts. While temporal coherence is appreciated most by viewing the corresponding videos, observations can be made with respect to Fig. 7. For example, notice the more consistent disparity estimates recovered for the low texture walls and the chair via stequel matching, the lack of sudden, high variation, seen both with **noST-local** and **noST-global**, and the more accurate outlines of the teddy bear, the head and the hat suspended above.

3.4 *Rover* sequence

The fourth 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.

For this case, prior to processing with the stereo algorithms, the sequence was stabilized in software to compensate for the extremely jerky camera motion: Stabilization operated by warping neighboring frames to reference frames throughout the video according to affine transformations recovered via a parametric motion estimator [4]. For presentation, however, results are shown with respect to the original (unstabilized) video.

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.

3.5 Motion estimation

To quantify the performance of the described 3D motion estimator, (29), a third lab dataset, *Lab 3*, is introduced; example frames are shown in Figure 9. The scene is composed of two vertically oriented, planar, textured rectangles that initially are frontoparallel with respect the camera. The left rectangle is relatively closer to the camera and rotates about the vertical. The right rectangle is relatively further from the camera and rotates about its base on an axis parallel to the optical axis. The cameras also move forward toward the rectangles parallel to the optical axis. For disparity groundtruth, the same methodology used for *Lab 1* and *Lab 2* is employed. To acquired 3D scene flow groundtruth, fiducial markers have been placed in the scene for reliable subpixel tracking (ARTag package was used for this purpose [16]). The markers are localized in successive frames, planes are fit and 3D motion in disparity space is recovered using an extant robust estimator [11] to yield dense disparity and flow maps within the surfaces.

Results of applying the described 3D motion estimator to the *Lab 3* dataset are shown in Fig. 10. Median angular

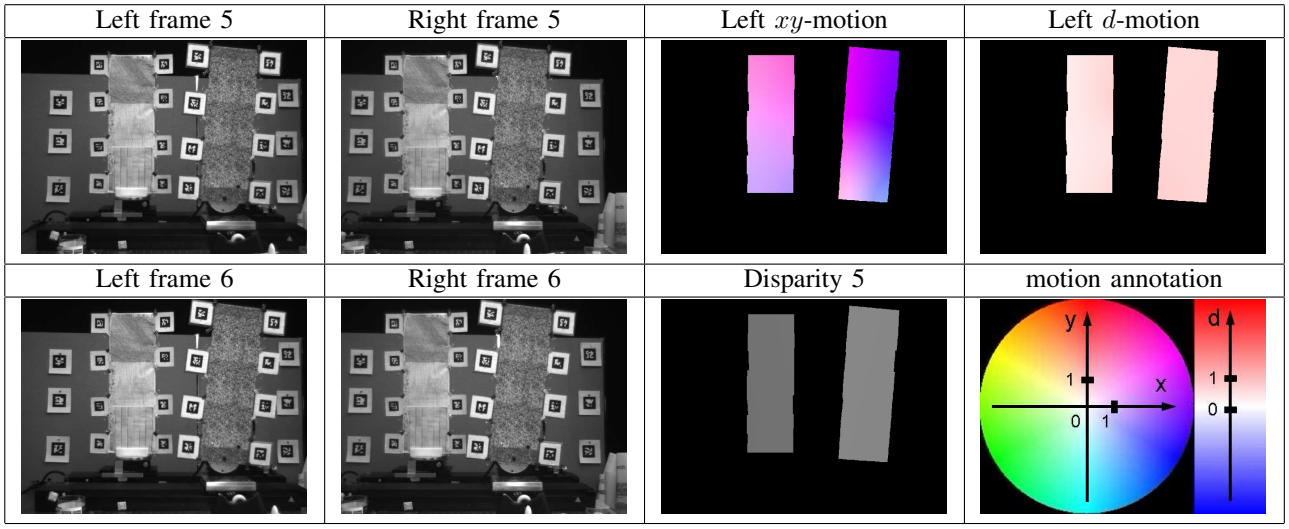


Fig. 9. *Lab 3* dataset with ground truth disparity and motion. Example motion estimation results for the consecutive pair of stereo frames $(x \times y \times t = 640 \times 480 \times 7)$. Left half of the figure shows the original intensity images for time consecutive frames, while right half show disparity and colour-coded flow components and the annotation chart associated with them. See accompanying videos at [44].

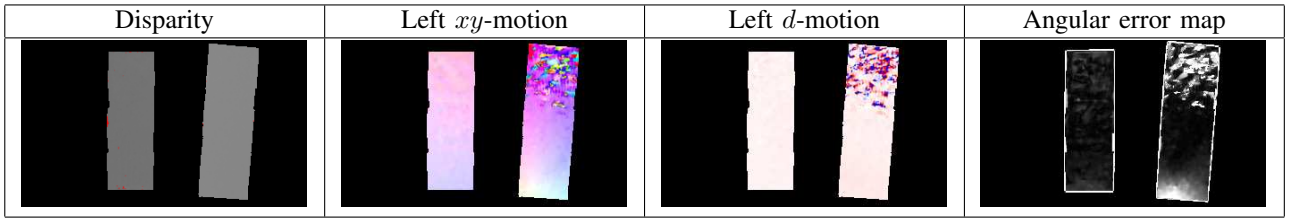


Fig. 10. Example motion estimation results for the middle frame of the *Lab 3* dataset for the **ST-local**. From left to right: recovered disparity, recovered xy-component of 3D motion; d-component of 3D motion; angular error map (black value corresponds to 0 and white corresponds to 90 or more degrees).

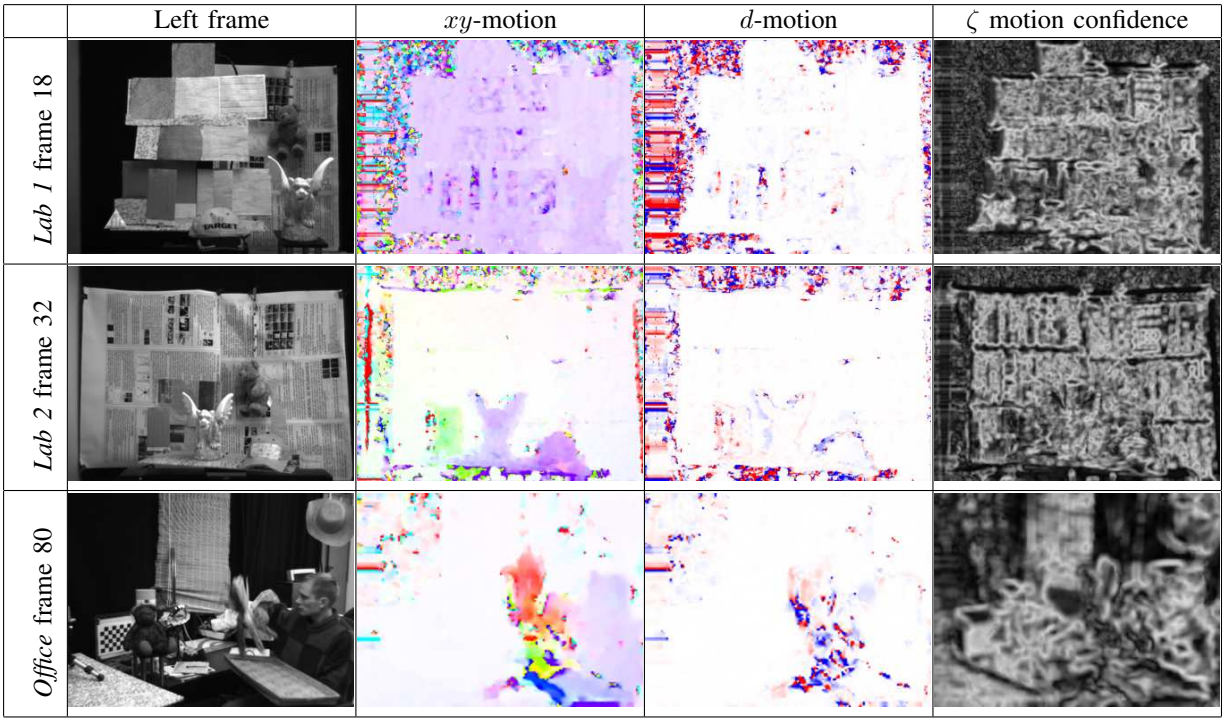


Fig. 11. Example motion estimation results for *Lab 1*, *Lab 2* and *Office* datasets (top and bottom rows, resp.). From left to right: left view, xy-component of 3D motion, d-component of 3D-motion map, ζ confidence map (brighter values corresponds to increased confidence). See accompanying videos at [44].

error between recovered and groundtruth 3D motion vectors across the entire dataset was 4.03 degrees. The angular error maps show that accuracy for the left surface where motion is about the vertical plus looming is especially good; whereas, limitations in capturing large motions parallel to the imaging plane are documented in the right surface. In particular, the error is seen to increase mostly as one moves toward the upper portions of the surface where the motion is largest due to the lever-arm nature of the set-up; use of a coarse-to-fine refinement scheme should increase the upper end of the range of motion magnitudes that can be recovered and is a possible direction for future research. Similarly, the colour coded flow plots show that the recovered estimates mostly are appropriately smooth, with breakdown in smoothness also occurring at larger magnitude motions.

Figure 11 shows the recovered motion for sample frames of the *Lab 1*, *Lab 2* and *Office* sequences. The flow vector confidence measure, ζ, (30) is also displayed. The recovered motion fields are qualitatively correct and appear quite smooth considering that no explicit optimization over flow vectors has been attempted. For *Lab 1* the light purple colour depicted as xy-motion reasonably accurately reflects the dominantly rightward motion; whereas, the light pink in d-motion corresponds to the camera moving forward. *Lab 2* results capture the rotation of the platform where the cap and the box are instantaneously headed in opposite horizontal directions (green and purple colours), because they are on different sides of the platform rotation axis; meanwhile the background is characterized with very light pink in d-motion map signalling camera moving forward. Finally, the *Office* sequences captures the motion of the person in the lower right of the image as moving rightward (light purple in xy-motion) and back (dark purple in d-motion), while picking up the bunch of corn from the tray (upward xy-motion coloured in red).

It also is illustrated that the confidence measure, ζ, reports reasonable values, e.g., highest in areas with enough image texture to yield adequate structure in Q, (28), to support motion estimates and low in untextured regions, such as the black backgrounds.

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 [44].

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. By necessarily committing to local flow vectors, especially when data is insufficient for such interpretation, optical flow yields unreliable temporal aggregation; in contrast, stequels more completely characterize whatever spatiotemporal structure is present and make it available for appropriate matching. 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 [31]). Stequels, on the other hand, are directly applicably to standard 2D MRF graphs and their successful performance has been documented in this paper.

Beyond their efficacy in establishing binocular correspondence for disparity estimation, stequels have been shown to provide the basis for estimation of 3D scene flow. In particular, matched stequels allow for direct recovery of 3D flow without the need for explicit, independent left and right image flow estimation. Again, the efficacy of this approach has been demonstrated empirically.

In conclusion, a computationally tractable and simple solution to spatiotemporal stereo and scene flow estimation has been presented, which proved to be very reliable, versatile and robust in practice. Significantly, the described research is the first to consider stequel matching for such purposes and various extension can be considered, e.g., exploiting the spatiotemporal profile for explicit non-Lambertian and multi-layer matching.

Acknowledgments

This work was supported by a CRD grant to R. Wildes, as funded by NSERC and MDA Space Missions.

References

[1] E. H. Adelson and J. R. Bergen. Spatiotemporal energy models for the perception of motion. *JOSA*, 2(2):284–299, 1985.
[2] S. Baker and I. Matthews. Lucas-Kanade 20 years on: A unifying framework. *IJCV*, 56(3):221–255, 2004.
[3] J. Bigun. *Vision with Direction*. Springer, 1998.
[4] M. J. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *CVIU*, 61(1):75–104, 1996.
[5] K. Cannons and R. P. Wildes. Visual tracking using pixelwise spatiotemporal oriented energy representation. In *ECCV*, pages 511–524, 2010.
[6] O. Chomat, J. Martin, and J. Crowley. A probabilistic sensor for the perception oand the recognition of activities. In *ECCV*, volume 1, pages 487–503, 2000.
[7] J. Davis, R. Ramamoorthi, and S. Rusinkiewicz. Spacetime stereo: A unifying framework for depth from triangulation. *TPAMI*, 27(2):296–302, 2005.
[8] D. Demirdjian and T. Darrell. Using multiple-hypothesis disparity maps and image velocity for 3D motion estimation. *IJCV*, 47:219–228, 2002.
[9] K. Derpanis, M. Sizintsev, K. Cannons, and R. Wildes. Efficient action spotting based on a spacetime oriented structure representation. In *CVPR*, 2010.

[10] K. Derpanis and R. Wildes. Dynamic texture recognition based on distributions of spacetime oriented structure. In *CVPR*, 2010.
[11] K. G. Derpanis and P. Chang. Closed-form linear solution to motion estimation in disparity space. In *IVS*, 2006.
[12] K. G. Derpanis and J. Gryn. Three-dimensional n-th derivative of Gaussian separable steerable filters. In *ICIP*, volume 3, pages 553–556, 2005.
[13] K. G. Derpanis and R. P. Wildes. Early spatiotemporal grouping with a distributed oriented energy representation. In *CVPR*, 2009.
[14] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *PETS*, 2005.
[15] M. Fahle and T. Poggio. Visual hyperacuity: Spatiotemporal interpolation in human vision. *Proceedings of the Royal Society of London, Series B, Biological Scienece*, 213(1193):451–477, 1981.
[16] M. Fiala. ARTag, a fiducial marker system using digital techniques. In *CVPR*, volume 2, pages 590–596, 2005.
[17] U. Franke, C. Rabe, H. Badino, and S. Gehrig. 6D-vision: Fusion of stereo and motion for robust environment perception. In *DAGM*, pages 216–223, 2005.
[18] W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *TPAMI*, 13(9):891–906, 1991.
[19] M. Gong. Enforcing temporal consistency in real-time stereo estimation. In *ECCV*, pages 564–577, 2006.
[20] G. Granlund and H. Knutsson. *Signal Processing for Computer Vision*. Kluwer, 1995.
[21] K. J. Hanna and N. E. Okamoto. Combining stereo and motion analysis for direct estimation of scene structure. In *ICCV*, pages 357–365, 1993.
[22] D. Heeger. A model for the extraction of image flow. *JOSA A*, 4:1455–1471, 1997.
[23] F. Huguet and F. Devernay. A variational method for scene flow estimation from stereo sequences. In *ICCV*, pages 1–7, 2007.
[24] M. Isard and J. MacCormick. Dense motion and disparity estimation via loopy belief propagation. *ACCV*, pages 32–41, 2006.
[25] M. Jenkin and J. K. Tsotsos. Applying temporal constraints to the dynamic stereo problem. *CVGIP*, 33:16–32, 1986.
[26] 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.
[27] A. Klaser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3D-gradients. In *BMVC*, 2008.
[28] V. Kolmogorov and R. Zabih. Computing visual correspondence with occlusions using graph cuts. In *ICCV*, pages 508–515, 2001.
[29] G. A. Korn and T. M. Korn. *Mathematical Handbook for Scientists and Engineers*. McGraw-Hill Companies, 2 edition, 1976.
[30] K. N. Kutulakos and S. M. Seitz. A theory of spape by space carving. *IJCV*, 38(3):199–218, 2000.
[31] E. S. Larsen, P. Mordohai, M. Pollefeys, and H. Fuchs. Temporally consistent reconstruction from multiple video streams. In *ICCV*, pages 1–8, 2007.
[32] C. Leung, B. Appleton, B. C. Lovell, and C. Sun. An energy minimisation approach to stereo-temporal dense reconstruction. In *ICPR*, pages 72–75, 2004.
[33] S. Malassiotis and M. G. Strintzis. Model-based joint motion and structure estimation from stereo images. *CVIU*, 65(1):79–94, 1997.
[34] 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.
[35] G. Medioni, C.-K. Tang, and M.-S. Lee. Tensor voting: Theory and applications. In *Proc. 12th Congres Francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence Artificielle*, 2000.
[36] J. Neumann and Y. Aloimonos. Spatio-temporal stereo using multi-resolution subdivision surfaces. *IJCV*, 47(1-3):181–193, 2002.
[37] Point Grey Research. http://www.ptgrey.com, 2010.
[38] J.-P. Pons, R. Keriven, O. Faugeras, and G. Hermosillo. Variational stereo & 3D scene flow estimation with statistical similarity measures. In *ICCV*, pages 597–602, 2003.
[39] W. Richards. Structure from stereo and motion. *JOSA*, 2:343–349, 1985.
[40] H. Scharr and R. Kusters. A linear model for simultaneous estimation of 3D motion and depth. In *WVM*, pages 220–225, 2002.
[41] H. Scharr and T. Schuchert. Simultaneous motion, depth and slope estimation with a camera-grid. In *Vision, Modelling and Visualization*, pages 81–88, 2006.
[42] D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. In *CVPR*, volume 1, pages 195–202, 2003.
[43] E. Shechtman and M. Irani. Space-time behavior-based correlation - or - how to tell if two underlying motion fields are similar without computing them? *TPAMI*, 29(11):2045–2056, 2007.
[44] M. Sizintsev. http://www.cse.yorku.ca/vision/research/spatiotemporal-stereo-stequel.shtml.
[45] M. Sizintsev and R. P. Wildes. Spatiotemporal stereo via spatiotemporal quadric element (stequel) matching. Technical Report CS-2008-04, York University, 2008.
[46] M. Sizintsev and R. P. Wildes. Spatiotemporal stereo via spatiotemporal quadric element (stequel) matching. In *CVPR*, 2009.
[47] M. Sizintsev and R. P. Wildes. Coarse-to-fine stereo vision with accurate 3D boundaries. *IVC*, 28(3):352–366, 2010.
[48] G. P. Stein and A. Shashua. Direct estimation of motion and extended scene structure from a moving stereo rig. *CVPR*, 1:211–218, 1998.
[49] G. Strang. *Introduction to applied mathematics*. Wellesley-Cambridge Press, 1986.
[50] C. Strecha and L. van Gool. Motion-stereo integration for depth estimation. In *ECCV*, pages 170–185, 2002.
[51] G. Sudhir, S. Baneerjee, K. K. Biswas, and R. Bahl. Cooperative integration of stereopsis and optic flow computation. *JOSA-A*, 12(12):2564–2572, 1995.
[52] S. Vedula, S. Baker, S. M. Seitz, and T. Kanade. Shape and motion carving in 6D. In *CVPR*, pages 2592–2598, 2000.
[53] A. M. Waxman and J. H. Duncan. Binocular image flows: Steps toward stereo-motion fusion. *TPAMI*, 8(6):715–729, 1986.
[54] J. Weng, P. Cohen, and N. Rebibo. Motion and structure estimation from stereo image sequences. *RA*, 8:362–382, 1992.
[55] O. Williams, M. Isard, and J. MacCormick. Estimating disparity and occlusions in stereo video sequences. *CVPR*, 1:250–257, 2005.
[56] A. Zaharescu and R. P. Wildes. Anomalous behaviour detection using spatiotemporal oriented energies, subset inclusion histogram comparison and event-driven processing. In *ECCV*, 2010.
[57] L. Zhang, B. Curless, and S. M. Seitz. Spacetime stereo: Shape recovery for dynamic scenes. In *CVPR*, pages 367–374, 2003.
[58] Y. Zhang and C. Kambhamettu. On 3D scene flow and structure estimation. In *CVPR*, volume 2, pages 778–785, 2001.
[59] 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.

Mikhail Sizintsev (Student Member, IEEE) received the BSc (Honours) and MSc degrees in computer science from York University, Toronto, Canada in 2004 and 2006, respectively. Currently, he is a PhD candidate at York University in the Department of Computer Science and Engineering. He spent the summer 2009 at Sarnoff Corporation in Princeton, New Jersey as an intern developing GPU-based stereo systems for augmented reality applications. His major areas of research are stereo and motion with specific emphasis in spatiotemporal processing and analysis.

Richard Wildes (Member, IEEE) received the PhD degree from the Massachusetts Institute of Technology in 1989. Subsequently, he joined Sarnoff Corporation in Princeton, New Jersey, as a Member of the Technical Staff in the Vision Technologies Group, where he remained until 2001. In 2001, he joined the Department of Computer Science and Engineering at York University, Toronto, where he is an Associate Professor and a member of the Centre for Vision Research. Honours include receiving a Sarnoff Corporation Technical Achievement Award, the IEEE D.G. Fink Prize Paper Award for his Proceedings of the IEEE publication “Iris recognition: An emerging biometric technology” and twice giving invited presentations to the US National Academy of Sciences. His main areas of research interest are computational vision, as well as allied aspects of image processing, robotics and artificial intelligence.