

Spacetime Stereo and 3D Flow via Binocular Spatiotemporal Orientation Analysis

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

Abstract—This paper presents a novel approach to recovering estimates of 3D structure and motion of a dynamic scene from a sequence of binocular stereo images. The approach is based on matching spatiotemporal orientation distributions between left and right temporal image streams, which encapsulates both local spatial and temporal structure for disparity estimation. By capturing spatial and temporal structure in this unified fashion, both sources of information combine to yield disparity estimates that are naturally temporal coherent, while helping to resolve matches that might be ambiguous when either source is considered alone. Further, by allowing subsets of the orientation measurements to support different disparity estimates, an approach to recovering multilayer disparity from spacetime stereo is realized. Similarly, the matched distributions allow for direct recovery of dense, robust estimates of 3D scene flow. The approach has been implemented with real-time performance on commodity GPUs using OpenCL. Empirical evaluation shows that the proposed approach yields qualitatively and quantitatively superior estimates in comparison to various alternative approaches, including the ability to provide accurate multilayer estimates in the presence of (semi)transparent and specular surfaces.

Index Terms—Stereo, motion, spacetime, spatiotemporal oriented energy, scene flow, multilayer reconstruction, transparency, specular.



1 INTRODUCTION

1.1 Motivation

THE goal of traditional binocular stereo is, given a pair of spatially separated 2D projections of a scene, recover the unknown third dimension of depth. This seemingly straightforward, versatile and essentially passive method of depth acquisition has been researched and matured for decades [45], [8]. Significantly, many applications acquire imagery over time and thereby allow for incorporation of the temporal dimension into processing. This additional information has the potential to resolve stereo matches that might be ambiguous when only instantaneous binocular views are considered. Further, availability of such information can help make 3D structure estimates temporally coherent and consistent with scene dynamics. Moreover, access to the temporal dimension supports recovery of 3D scene flow.

In response to these observations, the present work proposes a novel approach to spacetime stereo that relies on representing temporal image streams in terms of a distribution of 3D oriented energy measurements in visual spacetime, (x, y, t) . These energies are computed via application of a bank of spatiotemporal filters tuned to different orientations and applied separately to the left and right image streams for subsequent matching. Spacetime oriented analysis effectively captures both spatial and temporal

structure in a unified fashion, which allows their combination to drive matching for resolution of situations that might be ambiguous when either source is considered alone and increased temporal continuity. Further, by allowing subsets of the orientation measurements to support different estimates, a natural approach to multilayer disparity recovery arises that yields accurate results in the presence of depth discontinuities, (semi)transparency and specular reflections. Additionally, since spatiotemporal filter responses naturally encode scene dynamics, a direct method for recovery of 3D flow from orientation responses is derived.

1.2 Related research

Various attempts have been made to understand how the availability of temporal information can enhance binocular stereo. Some approaches smooth binocularly derived disparity estimates across consecutive temporal instants along optical flow directions [6] or along the temporal axis subject to change detection and background modeling [32]. Similarly, binocularly recovered surface mesh models have been smoothed across time by tracking [35], [38]. Other approaches reinforce disparity hypotheses by propagating correlation scores from the previous frame using optical flow [20]. A variety of other methods consider temporal information by extending a regularizing spatial MRF to include time and thereby allow for smoothing along the temporal direction, variously respecting flow displacements [31], either accounting for detected change [57] or not [34].

Other work has addressed simultaneous structure and motion estimation that combines intraframe (spatial) and interframe (temporal) image pairwise constraints in a wide variety of fashions. The formulations range from explicit

• *M. Sizintsev is with SRI International Sarnoff, Princeton, New Jersey, USA and R.P. Wildes is 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>*

stereo matching to recover depth for subsequent depth-flow estimation [52], [41], [18], [56], [26] to joint optimization formulations that simultaneously solve for structure and motion [53], [64], [27], [25], [55]. Typically, 3D flow estimation is the primary objective of these approaches, as they use availability of stereo information primarily to deduce the depth motion component and rely on standard binocular correspondence procedures for depth estimation.

Still other research combined binocular stereo and motion processing via explicit application of the brightness constancy constraint equation [23] across both modalities. For example, research has considered infinitesimal motion and stereo disparity estimation via a single brightness constancy formulation [43], [44]. Research along these lines also has made use of direct methods for integrated recovery of 3D scene structure and egomotion [22], [50], [36].

Another strand of research has been more generally concerned with recovering consistent depth maps across temporal image streams. For example, a bundle adjustment optimization approach for consistent depth map recovery across a monocular video sequence was proposed [62], which subsequently was extended to recover consistent depth maps across multiple synchronized video streams [59], [60], [28]. Other work also has considered recovery of consistent depth maps from multiple stationary video cameras [33]. While this body of research shares the current concern of depth map temporal consistency, it does not specifically address estimation from binocular video nor does it encompass other matters of concern (e.g., match disambiguation, multilayer estimates, scene flow estimation).

The proposed approach explicitly combines spatial and temporal support in stereo matching and thereby is most closely related to other research with similar concerns. One such method for spacetemporal stereo was initially developed in conjunction with temporally varying structured lighting [11]. Other work generalized this approach to model temporal disparity change [63]. Still other work extends the notion of spatially adaptive aggregation to include the temporal dimension [42]. Most closely related to the proposed approach is previous work by the authors that used measurements of spatiotemporal orientation as the basis for stereo matching [49]. That work encapsulated spacetemporal orientation in the spatiotemporal quadric or stequel (also referred to as the orientation tensor and covariance matrix [21], [4]) and was shown to yield disparity estimates with some degree of temporal coherence and ability to resolve otherwise ambiguous matches. However, representation in terms of the stequel fundamentally limits the ability to characterize the presence of multiple orientations at a point (as all are collapsed to a single quadric) that might further help distinguish matches, especially in situations involving multilayer surfaces (e.g., transparency) and near surface discontinuities. In contrast, the current approach makes more complete use of spacetemporal orientation measurements to allow for better resolution of difficult matching situations, including ability to resolve multilayer surfaces.

A major component of the proposed approach is the representation of imagery in terms of a distribution of

spatiotemporal oriented energy measurements. While previous research has exploited such measurements toward a variety of ends, e.g., optical flow recovery [2], dynamic texture analysis [16], tracking [9], video anomaly detection [61] and activity recognition [10], [13], it appears that no previous work has applied this approach directly to spacetemporal stereo. Other previous work made use of purely spatial orientation measurements in stereo matching [29], but it did not consider the temporal dimension.

1.3 Contributions

In the light of previous research, the outstanding contributions of the proposed approach are as follows. First, a novel method for spacetemporal disparity estimation is proposed based on direct matching of distributions of image spacetemporal orientation measurements. Second, the first approach to recovering multilayer disparity estimates from spacetemporal stereo processing is proposed. It is shown to allow for recovery of multiple layers in the presence of (semi)transparent and specularly reflecting surfaces. Interestingly, previous work in multilayer surface recovery from multiple images largely considers stereo (e.g., [46], [7], [54]) and motion (e.g., [3], [5]) only independently. Even previous work that combined multihypothesis disparity and optical flow for recovery of 3D motion estimates made use of purely binocular stereo considerations in its disparity estimation [12]. Third, a novel approach for direct estimation of dense 3D scene flow from binocular stereo-matched spatiotemporal orientation primitives is presented. Excepting previous work by the authors [49], this approach is the first to make use of binocularly matched spacetemporal orientation measurements for scene flow estimation and thereby enables recovery without the need for explicit temporal image correspondences. Moreover, it does so without the need to construct an intermediate stequel representation and thereby provides a more direct estimation approach compared to [49]. Fourth, the approach is realized in local and global stereo matchers with real-time GPU-based performance for the local version. Fifth, the developed implementations have been subject to extensive qualitative and quantitative empirical evaluation. A preliminary version of this research has appeared previously [48].

2 TECHNICAL APPROACH

2.1 Background

2.1.1 Interpretation of spatiotemporal orientation

Local oriented measurements in image spacetemporal, (x, y, t) , have visual significance. For example, orientations parallel to the image plane capture the spatial pattern of observed surfaces (e.g., texture); whereas, orientations that extend into the temporal dimensions capture dynamics (e.g., motion). The major proposition of this paper is that exploiting *spatiotemporal orientation energy distributions* (STE's) as primitive dynamic scene descriptors can provide a useful basis for spacetemporal stereo matching.

Figure 1 provides an illustrative example. Depicted is a dynamically rich scene containing moving people in

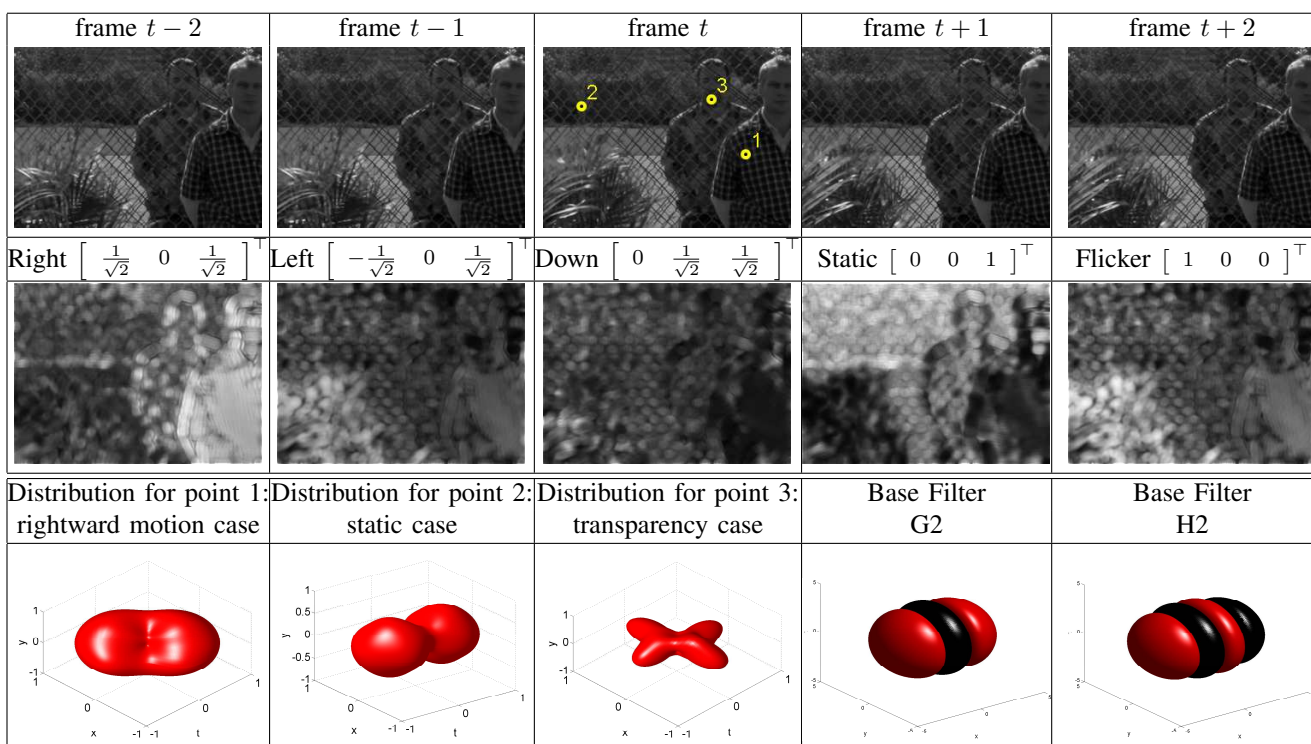


Fig. 1. Example of Spatiotemporal Volume and Pointwise Orientation Distributions. Top row: consecutive frames from a sample volume (originally presented in [14]). In the depicted scene, two persons move to the right, one in front and one behind a chain-link fence, rapidly moving foliage is in the lower left and the remainder is static. Middle row: response to particular spacetime orientations: rightward motion ($\hat{\mathbf{w}} = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}^\top$), leftward motion ($\hat{\mathbf{w}} = \begin{bmatrix} -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}^\top$), downward motion ($\hat{\mathbf{w}} = \begin{bmatrix} 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}^\top$), static ($\hat{\mathbf{w}} = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^\top$), and flicker ($\hat{\mathbf{w}} = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^\top$). Bottom row: spherical oriented energy distributions for points marked at the “frame t ” image, as well as the base steerable filter quadrature pairs used to extract energies.

front of and behind the pseudo-transparent fence as well as flickering foliage in the lower left. The top row of the figure shows 5 time-consecutive frames. This example illustrates important points to motivate the choice of matching primitive. First, as shown in the second row, different orientations capture different meaningful aspects of the scene structure and dynamics. Those that capture purely spatial static structure, e.g., $\hat{\mathbf{w}} = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^\top$, characterize surface texture for spatially driven matching; whereas, those that extend along spatiotemporal diagonals, e.g., $\hat{\mathbf{w}} = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}^\top$, capture motion to allow dynamics to constrain matching. Further, spatiotemporal orientations can capture flicker, e.g., $\hat{\mathbf{w}} = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^\top$, where temporal coherence breaks down. Second, as shown in the third row, by combining a sampling of oriented energy measurements along multiple directions at each point, a distribution is derived that can serve as a feature vector for subsequent matching. Both the static (point 1) and moving (point 2) samples have unique distributions that jointly characterize their spatial texture and dynamics. Further, the semitransparent point 3 shows a strongly bimodal distribution with potential for allowing multilayer disparity estimates, as different modes can support different disparities. Note that orientation distributions are always

symmetric with respect to the origin because orientations are taken as unsigned, i.e. $\hat{\mathbf{w}} \equiv -\hat{\mathbf{w}}$.

2.1.2 Measuring local spatiotemporal orientation

To exploit spatiotemporal orientation in binocular correspondence, one must commit to a particular approach to make local measurements of 3D, (x, y, t) , orientation in image spacetime data. Here, it proves to be advantageous to make use of oriented energy measurements based on steerable filters [19], as it will be shown they are amenable to matching directly on their responses to image data. In particular, recall that an energy measurement at a particular orientation, $\hat{\mathbf{w}}_i$, and spacetime position, $\mathbf{x} = (x, y, t)^\top$, can be obtained as the quadrature response of filtering image data $I(\mathbf{x})$ with Gaussian derivative filters of order n , $G_n(\hat{\mathbf{w}}_i)$ and their Hilbert transforms $H_n(\hat{\mathbf{w}}_i)$ as

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

with $*$ denoting convolution. Sample kernels for the second derivative quadrature filter pair G_2H_2 oriented at a particular orientation are depicted in the last row of Fig. 1.

Significantly, most of the practical uses of energy filtering of the form (1) involve a normalization step to make responses invariant to multiplicative bias and bring response values to the uniform scale 0 to 1. The necessary operation

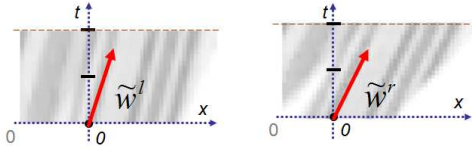


Fig. 2. The spatiotemporal orientation correspondence constraint, (3), describes the relationship between arbitrary orientations in correspondence, \hat{w}^l and \hat{w}^r , subject to binocular viewing of a slanted surface undergoing arbitrary motion in the world relative to the cameras. Depicted are the xt -slice views and one of several different orientation directions that might be considered across views.

is realized via pointwise division by the sum of the N local energy measurements at a point:

$$\hat{E}(\mathbf{x}; \hat{w}_i) = E(\mathbf{x}; \hat{w}_i) / \left(\sum_{j=1}^N E(\mathbf{x}; \hat{w}_j) \right). \quad (2)$$

Reasonably, N is taken as the number of orientations that span the space of orientations for the order of filtering that is employed. In the following, second-order, $n = 2$, Gaussians filters and their Hilbert transforms are used; so, $N = 10$ is required [19], with their orientations chosen to uniformly sample 3D orientation as the normals to the faces of an icosahedron with antipodal directions identified [40]. Essentially, the subsequent matching process operates on uniformly sampled distributions similar to the ones depicted in the last row of Fig. 1.

2.1.3 Spatiotemporal orientation correspondence

A prerequisite to the use of spatiotemporal orientation measurements for stereo matching is an analysis of how an arbitrary 3D world point that suffers an arbitrary displacement projects to related orientations in image spacetime, (x, y, t) , across a binocular pair. The essential result was presented in previous work by the authors [49] and is summarized here to provide necessary groundwork. Let unit vectors \hat{w}^l and \hat{w}^r (superscripts l and r denote left and right spacetimes, resp.) specify orientations about points that are in binocular correspondence, but otherwise arbitrary in visual spacetime as depicted in Fig. 2. These orientations are related as

$$\hat{w}^r = \frac{H\hat{w}^l}{\|H\hat{w}^l\|}, \quad \text{where } H = \begin{bmatrix} 1 + h_1 & h_2 & h_3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad (3)$$

with h_1 and h_2 capturing the motion independent change in local spatial orientations about corresponding points owing purely to the difference between binocular views of a (potentially) non-frontoparallel surface, while motion effects are captured by h_3 . Thus, the mapping between binocularly corresponding direction vectors, \hat{w}^l and \hat{w}^r , is governed by the parameters h_1, h_2, h_3 , with these parameters having the intuitive interpretation of accounting for the relative change of surface orientation across a binocular view as well as relative motion between the scene and sensor. A detailed

derivation is available elsewhere [49]. In the following, the basic relationship between binocularly corresponding image spacetime orientations, (3), will be referred to as the *spatiotemporal orientation correspondence constraint*.

2.2 Binocular spatiotemporal orientation error

With both the relationship between binocularly corresponding spatiotemporal orientations, (3), and a method for measuring local orientations, (2), in hand, an explicit stereo matching error can be developed.

The matching error is derived under the assumption that the pattern of the orientation distribution will vary between left and right views according to the binocular spatiotemporal orientation constraint, (3), but that it is otherwise appropriate to minimize the differences in the oriented filter responses. This approach amounts to a relaxed assumption of brightness constancy between views, as the filtered responses, (2), are robust to additive and multiplicative biases, which are discounted by the bandpass and normalized nature of the employed filters. In particular, the developed approach minimizes the sum of squared errors across all oriented energy measurements (2) as

$$\sum_{i=1}^N \epsilon_i^2(\mathbf{x}^l, \mathbf{x}^r) = \sum_{i=1}^N \left[\hat{E}^r(\mathbf{x}^r; \hat{w}_i) - \hat{E}^l(\mathbf{x}^l; \hat{w}_i) \right]^2, \quad (4)$$

which by (3) evaluates to

$$= \sum_{i=1}^N \left[\hat{E}^r \left(\mathbf{x}^r; \frac{H\hat{w}_i^l}{\|H\hat{w}_i^l\|} \right) - \hat{E}^l(\mathbf{x}^l; \hat{w}_i^l) \right]^2. \quad (5)$$

The error function, (5), is minimized by setting the corresponding gradient with respect to $\mathbf{h} = [h_1 \ h_2 \ h_3]^T$ to zero and subsequently solving for \mathbf{h} . Each error component ϵ_i^2 is a non-linear function of \mathbf{h} ; so, no closed form solution exists and numerical solutions will be noise sensitive owing to the high order in the variables of interest, \mathbf{h} . Instead, a solution is obtained via a first-order Taylor series expansion around $\mathbf{h}_0 = [0, 0, 0]^T$ to arrive at a simpler form for the error associated with each orientation \hat{w}_i as

$$\tilde{\epsilon}_i(\mathbf{x}^l, \mathbf{x}^r) = \epsilon_i(\mathbf{x}^l, \mathbf{x}^r; \mathbf{h}_0) + \nabla^T \epsilon_i(\mathbf{x}^l, \mathbf{x}^r; \mathbf{h}_0) \mathbf{h} \quad (6)$$

with $\epsilon_i(\mathbf{x}^l, \mathbf{x}^r; \mathbf{h}_0)$ the value of $\epsilon_i(\mathbf{x}^l, \mathbf{x}^r)$ at $\mathbf{h} = \mathbf{h}_0$.

2.3 Spatiotemporal orientation match cost

In this section, two approaches are presented for assigning a cost to matching points $\mathbf{x}^l = (x^l, y^l, t)^T$ and $\mathbf{x}^r = (x^l + d, y^l, t)^T$ across a binocular view according to disparity estimate, d , based on the error measure (6).

For the first approach, the linearized errors (6) for all orientations are combined into a system of linear equations

$$\mathbf{B}\mathbf{h} = \mathbf{b}, \quad (7)$$

where \mathbf{B} is an $N \times 3$ matrix, \mathbf{b} is an $N \times 1$ vector and $N = 10$ is the number of orientations measured, with

$$B_{i,m} = \frac{\partial \epsilon_i(\mathbf{x}^l, \mathbf{x}^r; \mathbf{h}_0)}{\partial h_m}, \quad (8)$$

and

$$b_i = \epsilon_i(\mathbf{x}^l, \mathbf{x}^r; \mathbf{h}_0), \quad (9)$$

so that each row of \mathbf{B} captures the error contribution of a particularly sampled direction $\hat{\mathbf{w}}_i$. A solution for \mathbf{h} is obtained by following standard linear algebraic manipulations [51] as $\mathbf{h} = (\mathbf{B}^\top \mathbf{B})^{-1} \mathbf{B}^\top \mathbf{b}$ with residual error

$$\tilde{\epsilon}^2 = \sum_{i=1}^N \tilde{\epsilon}_i^2(\mathbf{x}^l, \mathbf{x}^r) = \left(\mathbf{B} (\mathbf{B}^\top \mathbf{B})^{-1} \mathbf{B}^\top \mathbf{b} - \mathbf{b} \right)^2. \quad (10)$$

Thus, for any given disparity, d , the cost associated with matching \mathbf{x}^l with \mathbf{x}^r is taken as the residual, (10).

The second approach to assigning match cost on the basis of orientation measurements allows for multiple disparity estimates to be recovered at a point or over a spatial region. This approach relies on the fact that when multiple spatiotemporal orientations are superimposed or juxtaposed during image formation, their individual orientations persist in the composite imagery [15]. In particular, such combined orientation structure holds for a wide variety of naturally occurring phenomena, including both additive and multiplicative combinations to encompass, e.g., (semi)transparency, reflections and structure near boundaries of overlapped surfaces. Correspondingly, rather than combine all orientation measurements into a single system of equations in support of a single minimal cost disparity estimate, subsets of measurements can contribute to different estimates to enable multi-layer disparity estimation. Significantly, this approach allows for the recovery of multilayer estimates without the complications entailed in coordinated recovery of alpha-mattes (c.f., [58], [65]).

To realize these intuitions on multilayer disparity estimation, notice that for each matched point pair $\mathbf{x}^l, \mathbf{x}^r$ and orientation $\hat{\mathbf{w}}_i$ the Taylor expansion (6) yields a linear constraint on \mathbf{h}

$$\nabla^\top \epsilon_i(\mathbf{x}^l, \mathbf{x}^r; \mathbf{h}_0) \mathbf{h} = \epsilon_i(\mathbf{x}^l, \mathbf{x}^r; \mathbf{h}_0). \quad (11)$$

Geometrically, each such equation defines a plane in \mathbf{h} -space. Therefore, given multiple orientations aggregated over a region, the point in \mathbf{h} -space where the maximum number of planes intersect (and at least three in general position) defines the \mathbf{h} value that has the most support for a given disparity across the aggregation region. In the present implementation, a Hough transform [24], [17] is used to find the desired peak(s) in \mathbf{h} -space. To calculate the final match cost associated with any given peak, all points and orientations that contributed to a given peak are declared as inliers, denoted with *inl* subscripts, and used to calculate a residual error

$$\tilde{\epsilon}_{inl}^2 = \frac{1}{\#inl} \left(\mathbf{C} (\mathbf{C}^\top \mathbf{C})^{-1} \mathbf{C}^\top \mathbf{c} - \mathbf{c} \right)^2, \quad (12)$$

which has the same form as the match cost, (10), with

$$\mathbf{C}_{i,m} = \frac{\partial \epsilon_i(\mathbf{x}^l, \mathbf{x}^r; \mathbf{h}_0)}{\partial h_m}, c_i = \epsilon_i(\mathbf{x}^l, \mathbf{x}^r; \mathbf{h}_0), \quad (13)$$

and $i \in inl$, i.e., analogously to \mathbf{B} and \mathbf{b} in (8) and (9). Additional normalization by $\#inl$ is desired because the

number of inliers that contribute to the error, (12), generally varies for different disparity hypotheses, while corresponding error values must be directly comparable for effective subsequent local or global disparity optimization strategies. Once the inliers are determined and the corresponding final error is computed for all disparities to be considered, only those disparity hypotheses that have enough supporting evidence, i.e. the number of contributions is above a corresponding threshold, λ_{inl} , are accepted.

The first method for assigning cost to disparity estimates, (10), allows for recovery of only a single disparity estimate at a point. In that sense, it is analogous to the earlier approach to disparity estimation based on the spatiotemporal quadric element, [49]; however, the proposed approach makes more complete use of available orientation information by eschewing its collapse to the quadric. The second method, (12), allows for multilayer disparity estimates over a window of match aggregation by more fully exploiting the availability of multiple orientation measurements. In contrast, the earlier method [49] was not able to exploit this possibility because in collapsing all orientation information into the quadric it sacrificed the degrees of freedom that make multilayer disparity estimation possible.

2.4 Scene Flow Estimation

2.4.1 3D scene flow

To begin, consider a single (i.e. left or right) spacetime volume. In this case, the 3D, (x, y, t) , direction associated with a 2D, (x, y) , image flow, \mathbf{v}_2 , must correspond to a minimal energy across orientations, as brightness constancy assumes uniform intensity along the direction of flow [21], [4]. Thus, to solve for the appropriate direction, the basis set of oriented energy measurements, (1), could be steered to the direction that yields minimal energy response. In the context of 3D scene flow, \mathbf{v}_3 , recovery from corresponding left and right image spacetime directions, (x, y, t) , must perform steering as a joint optimization across both views.

The key challenge is to choose the appropriate parametric representation for the 3D orientation. Here, it is advisable to avoid use of unit vectors $\hat{\mathbf{w}}$, since explicit normalization would be required. Assuming rectified views, the epipolar constraint dictates that the y -component of the motion must be the same in left and right views, while the x and depth components are coupled according to (assuming the left image is the reference view):

$$x^r = x^l + d^l, y^r = y^l. \quad (14)$$

Having the latter in mind, the spacetime unit vector $\hat{\mathbf{w}}$ is parameterized by the angle pair (α, β) , where α is the angle between the t -axis projection of $\hat{\mathbf{w}}$ onto the yt plane and β is the angle between $\hat{\mathbf{w}}$ and the x -axis. Unit vector $\hat{\mathbf{w}} = [w_x \ w_y \ w_t]^\top$ now is represented in terms of (α, β) as

$$\hat{\mathbf{w}} = \begin{bmatrix} w_x \\ w_y \\ w_t \end{bmatrix} = \begin{bmatrix} \cos(\beta) \\ \sin(\alpha) \sin(\beta) \\ \cos(\alpha) \sin(\beta) \end{bmatrix} = \hat{w}(\alpha, \beta). \quad (15)$$

In the current context for compactness of notation, let the measurement of oriented energy, (1), for the left spacetime

volume, I^l , along direction $\hat{\mathbf{w}}^l$ at a particular location be given as $E^l(\hat{\mathbf{w}}^l)$ and analogously for the right spacetime volume. The particular parametrization of angles, (15), facilitates joint optimization of $E^l(\hat{\mathbf{w}}^l) = E^l(\hat{w}(\alpha^l, \beta^l))$ and $E^r(\hat{\mathbf{w}}^r) = E^r(\hat{w}(\alpha^r, \beta^r))$, as the epipolar constraint dictates that $\alpha^l = \alpha^r = \alpha$. Thus, the objective is to find the parameters $(\alpha, \beta_l, \beta_r)$ that minimize

$$\begin{aligned} E^{lr} &= E^l(\hat{\mathbf{w}}^l) + E^r(\hat{\mathbf{w}}^r) \\ &= G^l(\hat{\mathbf{w}}^l)^2 + H^l(\hat{\mathbf{w}}^l)^2 + G^r(\hat{\mathbf{w}}^r)^2 + H^r(\hat{\mathbf{w}}^r)^2. \end{aligned} \quad (16)$$

As suggested in the first paragraph of this section, the optimization is accomplished by simultaneously steering the left and right basis oriented energy responses, (1), so that the objective, (16), is minimized. Here, the minimization is carried out in a straightforward two step process. First, take as an initial estimate of $(\alpha, \beta^l, \beta^r)$ the pair of orientations that minimizes (16) across the set of orientations explicitly calculated for (1), i.e. the ten icosahedral defined directions. Second, apply Gauss-Newton minimization for incremental refinement. To formulate the Gauss-Newton increment, the current error contribution and its Jacobian must be specified. In the present context, these are given as

$$\begin{aligned} \mathbf{r}(\alpha, \beta^l, \beta^r) &= [G^l(\hat{\mathbf{w}}^l) \ H^l(\hat{\mathbf{w}}^l) \ G^r(\hat{\mathbf{w}}^r) \ H^r(\hat{\mathbf{w}}^r)]^\top, \\ \mathbf{J}(\alpha, \beta^l, \beta^r) &= (\Gamma \mathbf{r}^\top)^\top, \end{aligned} \quad (17)$$

where $\Gamma = \begin{bmatrix} \frac{\partial}{\partial \alpha} & \frac{\partial}{\partial \beta^l} & \frac{\partial}{\partial \beta^r} \end{bmatrix}^\top$.

Finally, once the solution for $(\alpha, \beta^l, \beta^r)$ is obtained, the unit directional vectors $\hat{\mathbf{w}}^l$ and $\hat{\mathbf{w}}^r$ are computed according to (15) from which the 3D disparity flow is computed as

$$\mathbf{v}_3 = \begin{bmatrix} \frac{w_x^l}{w_t^l} \\ \frac{w_y^l}{w_t^l} \\ \frac{w_x^r}{w_t^r} - \frac{w_x^l}{w_t^l} \end{bmatrix} = \begin{bmatrix} \sec(\alpha) \cot(\beta^l) \\ \tan(\alpha) \\ \sec(\alpha) (\cot(\beta^r) - \cot(\beta^l)) \end{bmatrix}. \quad (18)$$

Provided binocular camera calibration, actual 3D scene flow then can be recovered from the disparity flow analogously to recovery of 3D distance from disparity, as desired.

2.4.2 Uncertainty of flow estimation

The local quadratic surface approximation near the minimum of the objective function, E^{lr} , (16), is a good indication of uncertainty in the context of the employed Gauss-Newton method. Essentially, the Hessian $\mathcal{H}_{\mathcal{M}}$ of (16) derived analytically is capable of describing the behavior of the recovered solution. Here, the Hessian is a 3×3 matrix of second derivatives of E^{lr} w.r.t. α , β^l and β^r .

Specifically, let $\lambda_1 \geq \lambda_2 \geq \lambda_3$ be the eigenvalues of non-negative-definite $\mathcal{H}_{\mathcal{M}}$. The case of a well-defined (flow) vector corresponds to the condition $\lambda_3 \gg 0$. Moreover, any of the λ_i being close to zero indicates how underconstrained the flow is, e.g., instances of the aperture problem where only the normal flow can be recovered. Thus, the chosen measure of the recovered 3D flow confidence is

$$\zeta = \lambda_{min}, \quad (19)$$

where λ_{min} is the smallest eigenvalue of $\mathcal{H}_{\mathcal{M}}$.

3 EXPERIMENTAL EVALUATION

3.1 Algorithmic Instantiations

Implementations of the proposed approach accept synchronized and rectified binocular videos, I^l, I^r , as input, recover basis orientation measurement distributions, $\hat{E}^l(\hat{\mathbf{w}}_i^l)$, $\hat{E}^r(\hat{\mathbf{w}}_i^r)$ and then calculate the match cost for any given disparity, d , using one of the methods (10) or (12). The match cost has been embedded in both local and global stereo matchers, denoted **STE-local** and **STE-global**, resp., to illustrate the broad applicability of the approach. The local algorithm is an adaptive, coarse-to-fine block-matcher operating over Gaussian pyramids [47]. The global algorithm is a graph-cuts matcher [30]. These particular matchers were chosen because they have been used previously in realizing the stequel approach to spacetime stereo [49] and thereby allow for direct comparison. Further, given matched orientation distributions, estimates of 3D scene flow are obtained via the motion recovery procedure of Sec. 2.4.1.

The local matching approach makes use of the per orientation match cost, (12), to support recovery of multiple layer disparity estimates. In all cases spatial aggregation is 5×5 and inlier threshold $\lambda_{int} = 4$. The global method makes use of the across orientations match cost, (10), with no spatial aggregation to avoid non-trivial optimization involving multiple label association, which is beyond the scope of the current paper. The global method is thereby not capable of multilayer estimation. In preliminary investigation, the across orientation cost, (10), also was embedded in the local method; it was found that results were extremely similar to those shown here for single layer disparity estimates and are not given explicitly. For both implementations, subpixel estimation was performed as post-processing using a Lucas-Kanade type refinement [1] specialized to the proposed spatiotemporal match costs, (10) and (12), as done analogously in previous work [49].

The local algorithm, **STE-local** is well suited to parallel computation and has been implemented in OpenCL [39] to be independent of hardware vendor. For the results presented here, this implementation was executed on an nVidia GTX580 GPU at 16 fps for 640×480 video with 256 disparity levels, where execution speed scales linearly with image size given in total pixels. Here, use of coarse-to-fine processing within image pyramids allows for computational efficiency even in the presence of large disparities, as large search at full image resolution is not required: It is encompassed via small search range at upper levels of the image pyramid [47]. The global algorithm, **STE-global** has been realized in C++ for execution on standard CPUs.

To demonstrate the benefits of the proposed spatiotemporal matching, several alternative approaches are compared. First, comparison is made to conventional spatial-only matching using image intensity with normalized correlation match cost, as realized in both local adaptive, coarse-to-fine block [47] and global graph-cut [30] algorithms; these methods will be denoted **noST-local** and **noST-global**, resp. Second, comparison is made to the ancestor of the currently proposed approach, stequel-based matching, again with

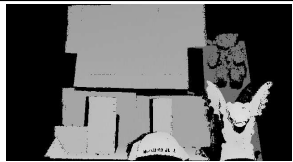
















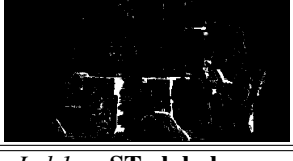
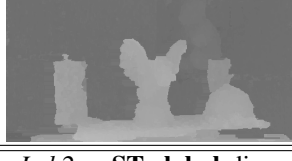



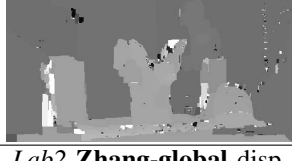



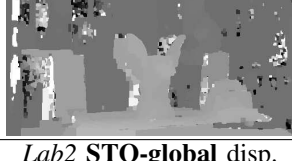

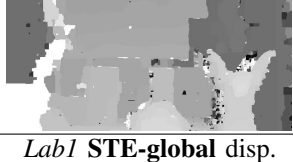

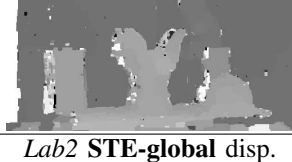





<i>Lab1</i> Disparity GT	<i>Lab1</i> Left frame	<i>Lab2</i> Disparity GT	<i>Lab2</i> Left frame
			
<i>Lab1</i> noST-local disp.	<i>Lab1</i> noST-local error	<i>Lab2</i> noST-local disp.	<i>Lab2</i> noST-local error
			
<i>Lab1</i> Zhang-local disp.	<i>Lab1</i> Zhang-local error	<i>Lab2</i> Zhang-local disp.	<i>Lab2</i> Zhang-local error
			
<i>Lab1</i> STQ-local disp.	<i>Lab1</i> STQ-local error	<i>Lab2</i> STQ-local disp.	<i>Lab2</i> STQ-local error
			
<i>Lab1</i> STE-local disp.	<i>Lab1</i> STE-local error	<i>Lab2</i> STE-local disp.	<i>Lab2</i> STE-local error
			
<i>Lab1</i> noST-global disp.	<i>Lab1</i> noST-global error	<i>Lab2</i> noST-global disp.	<i>Lab2</i> noST-global error
			
<i>Lab1</i> Zhang-global disp.	<i>Lab1</i> Zhang-global error	<i>Lab2</i> Zhang-global disp.	<i>Lab2</i> Zhang-global error
			
<i>Lab1</i> STQ-global disp.	<i>Lab1</i> STQ-global error	<i>Lab2</i> STQ-global disp.	<i>Lab2</i> STQ-global error
			
<i>Lab1</i> STE-global disp.	<i>Lab1</i> STE-global error	<i>Lab2</i> STE-global disp.	<i>Lab2</i> STE-global error
			

Fig. 3. Example input frames, groundtruth, recovered disparity and absolute difference error for *Lab1* and *Lab2*.

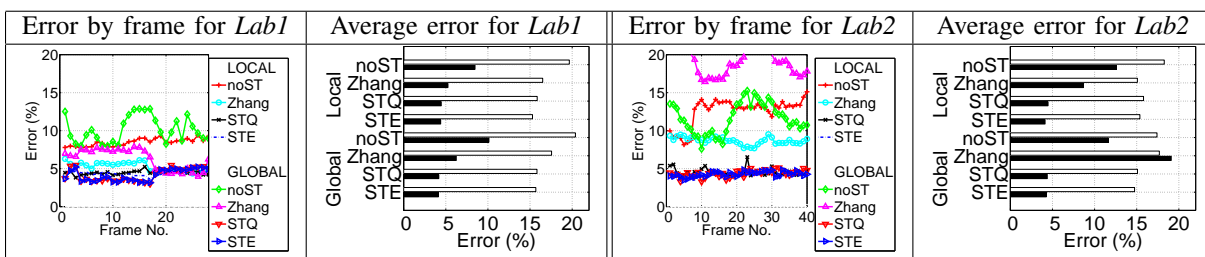


Fig. 4. Summary statistics for *Lab1* and *Lab2*. An error is taken as greater than 1 pixel discrepancy between recovered and groundtruth disparity. Bar plots show average error across entire sequences: White bars are for points within 5 pixels of a surface discontinuity; black bars show overall error. Error by frame plots show percentage of points in error overall for each frame separately.

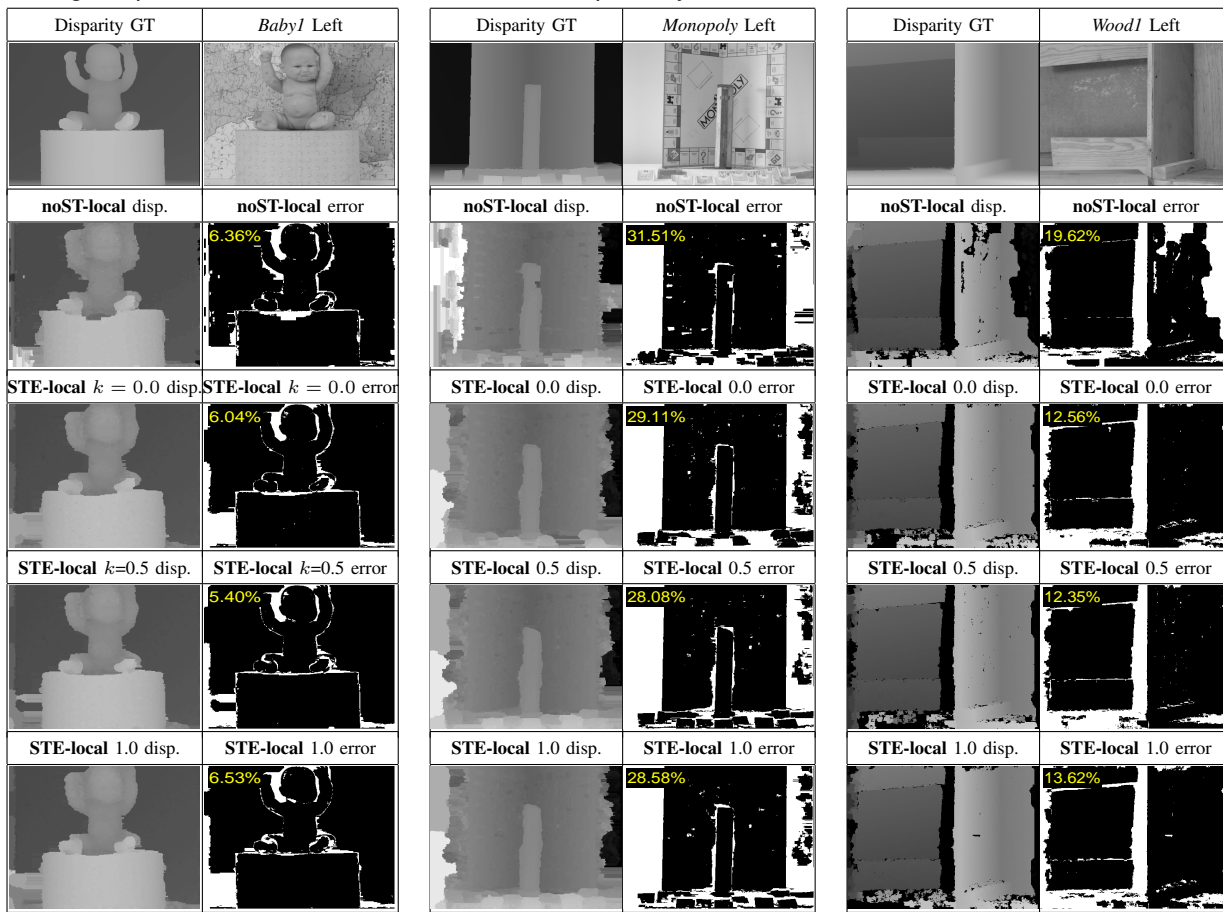


Fig. 5. Example input frames and recovered disparity maps for the modified *Middlebury* dataset [37]. For each example, original left, disparity ground truth, disparity and error maps for **noST-local** and **STE-local** algorithms are depicted. Error rates for occluded points with threshold of 2 disparity levels is given in upper right corner of the corresponding error map.

both local and global instantiations [49], denoted **STQ-local** and **STQ-global**, resp. Third, an alternative space-time stereo approach that uses image intensity matching with spatiotemporal oriented aggregation will be considered [63]. As with all others instances, this approach has been implemented within the same local [47] and global [30] matchers, denoted **Zhang-local** and **Zhang-global**, resp.

3.2 Lab datasets

In total, ten binocular video data sets are used as input. The first two are the *Lab1* and *Lab2* videos originally presented

elsewhere [49]. These sets are considered as they are natural image sequences with disparity groundtruth and have been used previously in comparison of spacetime stereo algorithms. Challenges present in these videos include weak, epipolar-aligned and camouflaging surface texture, complex 3D shapes (e.g., gargoyle and teddy bear) and a wide range of motions (vertical, horizontal and depth axis translations in *Lab 1*, depth axis translation and out-of-plane rotations of non-trivial magnitudes in *Lab 2*).

Example input image frames, groundtruth disparity, recovered disparity and summary performance statistics are

presented in Figs. 3 and 4. The results show that both local and global versions of **STE** and **STQ** perform better than the **noST** algorithms that eschew temporal information. Attention to regions involving weak and epipolar aligned texture (e.g., the planar regions of *Lab1*) show that the inclusion of temporal information helps to resolve purely spatial match ambiguities. Consideration of the relative smoothness of the error time series provides evidence of improved temporal coherence offered by **STE** and **STQ**.

In these tests, the improvement of **STE** relative to **STQ** arises in the vicinity of depth discontinuities. While, the error statistics in Fig. 4 suggest only marginal improvement, visual inspection of the error maps in Fig. 3 show that **STE** produces smallest error in particularly challenging situations, e.g., as it does best in resolving the narrow gap between the gargoyle wings. This is particularly the case for **STE-local**, where improved resolution of structure near 3D boundaries is expected, as its ability to capture multiple disparities allows a consensus to develop that accurately segregates the foreground and background depths without allowing one to contaminate the other. Interestingly, near surface discontinuities, it can happen that two disparities corresponding to the foreground and background within the aggregation window exceed the voting threshold, λ_v ; typically, however, either the foreground or background dominates the voting depending on the aggregation support and only the dominant surface is recovered. Moreover, for half-occlusion, the occluded point fails to yield consensus voting as no meaningful match is available.

Interestingly, spatiotemporal matching based directly on intensities, **Zhang**, did not show significant advantages even over purely spatial stereo, **noST**, and behaves noticeably worse than **STQ** and **STE**. Still, for *Lab1*, **Zhang** does help disambiguate matches in the camouflage (lower left) and epipolar-aligned texture regions relative to **noST**. Its performance on *Lab2* is particularly poor, especially in the fine-textured background regions, which can be explained by the zooming effect associated with in-depth motion that is not effectively captured by the simple temporal window shifts adopted in [63]. In particular, **Zhang** performs explicit temporal aggregation across shifted windows in time to form its binocular match support, which is only an accurate model when motion is translation parallel to the stereo baseline; in other cases the inapplicability of the model appears to lead to noisy disparity estimates. In contrast, spatiotemporal oriented energy distributions are pointwise measurements of the first-order intensity structure and explicit temporal aggregation is not performed during the **STE** matching procedure; hence, no such problem arises.

3.3 Middlebury dataset

To place the developed approach within the larger context of contemporary computer vision stereo vision research, it has been evaluated on samples from the standard Middlebury dataset [37]. Since the current algorithms operate on temporal streams of images, binocular frame pairs from the dataset have been extended to binocular videos via warping

with synthetic 2D flow fields, \mathbf{v} , generated according to

$$\mathbf{v}(x, y, t) = k|t|d(x, y, 0)\hat{\mathbf{u}}, \quad (20)$$

where k is a speed scale factor, $|t|$ is the absolute value of time, $d(x, y, 0)$ is the groundtruth disparity and $\hat{\mathbf{u}}$ is the direction of motion unit vector. This warping function is employed as it reasonably models the equations of the visual motion field for the case of scene translation orthogonal to the optical axis [23], e.g., resulting in visual motion parallel to the scene motion and inversely proportional to depth. For the present experiments, $t \in [-2, 2]$ is time relative to the original (central) frame, $\hat{\mathbf{u}} = [0 \ 1]^T$ is vertical to maximize the temporal information relative to the horizontal stereo baseline and k takes values 0.0, 0.5 and 1.0 to simulate zero, slow (within pixel) and faster (more than a pixel) motions, resp.

Selected test cases from the Middlebury dataset are *Baby1*, *Monopoly* and *Wood1*, as they tend to challenge many local binocular algorithms owing to their relatively impoverished texture and this challenge is one that the proposed spatiotemporal approach is argued to ameliorate by augmenting spatial information with temporal. Results are shown in Fig. 5 for **STE-local** as well as **noST-local** for the sake of comparison to purely spatial disparity estimation. It is seen that mere inclusion of temporal support during matching improves performance relative to reliance on purely spatial matching, as indicated in the decreased error of **STE-local** compared to **noST-local**. Moreover, when motion-based displacement complements that of disparity the benefit further increases, as shown by the case of $k = 0.5$. Finally, increasing the speed beyond a certain point yields diminished returns, as shown with even larger $k = 1.0$, especially near horizontal depth boundaries (e.g., *Baby1*) that are orthogonal to the direction of motion $\hat{\mathbf{u}}$. Here, temporal aliasing can ensue and also boundary artifacts are introduced by the image synthesis, which contaminate spatiotemporal filtering results. Similar effects show near the bottom of *Wood1*, as the originally imaged low texture surfaces are narrow along the direction of vertical motion and become dominated by synthesis artifacts that in turn corrupt the spatiotemporal filtering.

3.4 Skydiving dataset

To illustrate the performance of the developed approach in challenging real-world scenarios (albeit without groundtruth), Fig. 6 shows a binocular sequence of a group of skydivers during their descent. Challenges include the complicated motion patterns of the skydivers (both downward and spiral/left-right), their intricate spatial relationships (e.g., juxtapositions and occlusions) and the overall unconstrained nature of the acquisition. Here, comparison focusses on the relative performance of the local matchers **noST-local**, which does not employ temporal information, the alternative spatiotemporal matcher **Zhang-local** and the proposed **STE-local**. In general, it is seen that introduction of temporal information in **Zhang-local** and **STE-local** yields better temporal coherence in the disparity estimates in comparison to **noST-local**, including reasonable

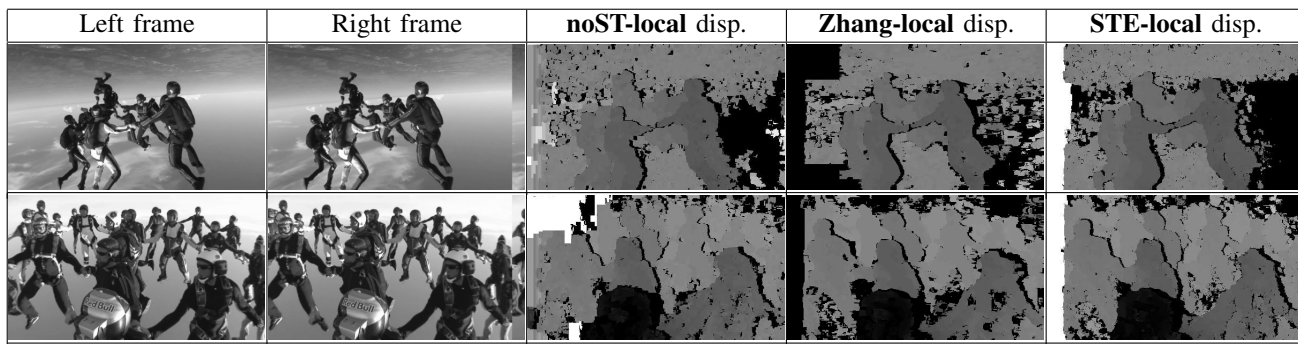


Fig. 6. Example input frames and disparity maps for the *Skydiving* dataset [42]. This dataset was rectified such that zero values correspond to close distance rather than infinity; thus, brighter values mean greater depth.

results in the poorly textured background sky. Furthermore, it appears that **STE-local** results provide generally the crispest 3D boundaries owing to the algorithm's ability to distinguish multilayer disparity relations.

3.5 Transparency dataset

An important distinguishing point of **STE-local** in comparison to all previous spatiotemporal stereo algorithms is the ability to deal with multilayer disparities at a point. The next test dataset, *Transparency*, illustrates the case of semitransparency. This real image sequence was captured by placing an acetate film in front of a background surface with each of the two surfaces covered by a different texture pattern such that the foreground is semitransparent while the background is opaque. One of the surfaces was set in horizontal motion and captured binocularly, see Fig. 7. Consideration of a single left/right frame pair makes it very difficult to recover the two disparity layers that are present; however, since the two surfaces are in relative motion, they create distinctive spacetime orientation patterns. The superposition of these two patterns are readily apparent in the illustrated xt -slices, where the vertical and diagonal orientations arise from the stationary background and translating foreground surfaces, resp. In essence, different orientations correspond to layers residing at different depths and certain orientations will be consistent with one layer or another. A plot of cost, (12), as a function of disparity vs. spatiotemporal orientation also is shown in the figure. It is apparent that the smallest errors, i.e. darker colors, are concentrated about two disparity values (approximately 120 and 175), which correspond to the foreground and background surfaces. Also shown is the distribution of votes accumulated by **STE-local** for different disparities across the entire sequence, which shows a strongly bimodal distribution. Finally, a perspective surface plot of the disparities recovered by **STE-local** for a particular frame pair is displayed that shows the presence of two disjoint layers. Note that **STE-local** only offers multiple disparities at a point, but not the explicit grouping of underlying layers, which is taken as later visual processing.

To underline the importance of spatiotemporal orientation in multilayer matching, an alternative multilayer matcher that works directly on single left/right frame pairs

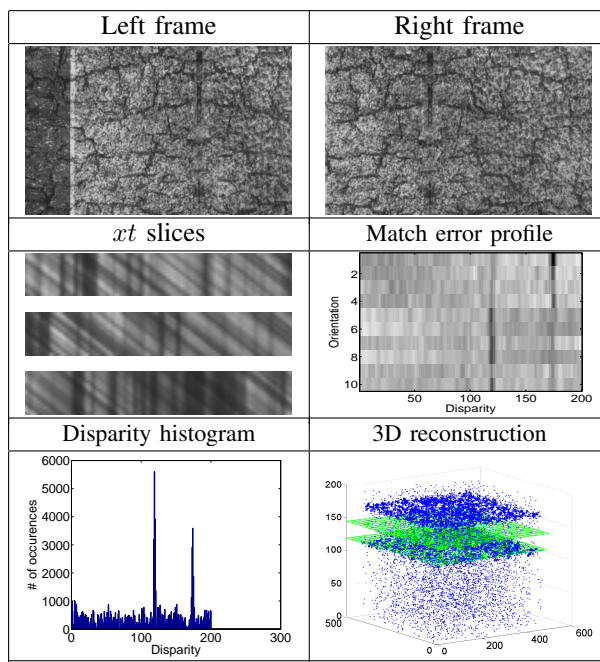


Fig. 7. Example input frames, spacetime slices and disparity estimation results for *Transparency*. See text for details.

also was applied to the *Transparency* data set. This matcher makes use of robust, parametric layer estimation [5] and was applied to the left and right frames that arise half way through the *Transparency* sequence. The results are plotted as green planes in the perspective plot of Fig. 7. It is seen that the background surface is reasonably recovered at disparity 120.75; however, the foreground surface is greatly underestimated at disparity 148.19 (correct disparities are 120 and 175, resp.). Apparently, the intensity mixtures that result from semi-transparency cannot be separated properly by robust application of brightness constancy, as employed by the alternative approach; whereas, the proposed approach based on explicit representation of multi-oriented intensity structure allows for success.

3.6 Lustre dataset

While the case of transparency is complicated and intriguing, specular reflections are more common in practice. Indeed, relatively few surfaces are purely matte, especially

in the man-made world. The dataset *Lustre* deals with the case of “binocular lustre” where a specular reflection is present in one of the two views and totally absent in the other. This sequence was acquired by having a well textured planar surface that is covered with a reflective coating rotate about the horizontal during binocular video capture. Across the sequence, an overhead light is strongly reflected in the left view, but not present in the right, see Fig. 8. Here, comparison focusses on the previous **STQ-local** and the newly proposed **STE-local** to underline the distinctions of the latter. The spatiotemporal stequel matcher **STQ-local** is able to reasonably capture the surface outline and some of the interior. However, the proposed **STE-local** algorithm achieves improved results as it can better capitalize on those components of the spatiotemporal orientation distributions that have reliable matches across views and ignore those that do not. In contrast, purely spatial matching, **noST-local** performs much poorer as it has no basis to overcome the incompatible intensity profiles that arise due to lustre.

3.7 Bino-spec dataset

The dataset, *Bino-spec*, deals with the case of binocular specularity, where a specular reflection is present in both views, but is displaced in mirror fashion relative to the underlying surface. This sequence was acquired by having a well textured, cylindrical cup with a shiny coating rotating about a vertical axis. Throughout the sequence, a window in the room is strongly reflected in both views, see Fig. 9. Pixel matcher **noST-local** is able to recover the cup outline, but fails to match correctly the interior portion due to its high reflectivity and the presence of superimposed disparities of the cup texture and specular reflection. At these points the algorithm recovers the surface, the reflection or some erroneous mixture. In contrast, **STE-local** is able to recover two disparity layers, as appropriate. The depicted “primary estimate” map shows the disparity at each point that received the top number of votes above λ_{int} , the “secondary estimate” map shows other disparities whose number of votes also surpassed λ_{int} , the majority of which are concentrated near the specularities on the cup and 3D boundaries (see above discussion of 3D boundaries). The top view 3D reconstruction shows the recovery of both the cup surface as well as the specularity properly placed behind the surface according to mirror reflection with respect to the convex surface. In comparison, when **STQ-local** was applied to this case the results were very similar to the primary estimate of **STE-local** (and therefore not shown in the interest of space); significantly, however, **STQ-local** is fundamentally incapable of recovering secondary estimates.

3.8 Motion estimation

To quantify the performance of the described 3D motion estimator, Sec. 2.4, an additional lab dataset originally introduced in [49], *Lab3*, is employed; example frames with groundtruth disparity and motion are shown in Fig. 10. The scene is composed of two vertically oriented, planar, textured rectangles that initially are frontoparallel with

respect to the camera. The left rectangle is relatively closer to the camera and rotates about the vertical. The right rectangle rotates about its base on an axis parallel to the optical axis. The cameras also move parallel to the optical axis, toward the rectangles.

Motion estimation results on the *Lab3* dataset are shown in Fig. 10. Median angular error between recovered and groundtruth 3D motion vectors across the entire dataset was 4.01 degrees. Qualitatively, it is seen that estimates are reasonably accurate and smooth for the smaller magnitude motions (along the depth axis, rotation about the vertical of the left surface, translation parallel to the image plane in the lower and middle portions of the right surface). Further, performance degrades reasonably smoothly with increased magnitude motion (e.g., notice the transition in error from the middle to upper portions of the right surface). Interestingly, the results presented here are quite comparable to those presented for the earlier 3D flow estimator associated with **STQ-local** [49], i.e. 4.03 degrees median error; however, the current algorithm provides more direct access to flow, as it operates directly on matched orientation distributions rather than abstracted stequels. In both cases, a potentially valuable direction for future research would be development of a coarse-to-fine refinement scheme to increase the motion capture range.

As a further comparison, 3D flow results were obtained for *Lab3* in a more standard fashion: Stereo disparity was recovered with **noST-local**, 2D flow was recovered separately for the left and right views with a robust Lucas-Kanade optical flow estimator [1] and 3D flow subsequently was inferred by appeal to the left/right flow relationship (14). In the following, this approach will be referred to as **noST-LK**. The resulting flow map in Fig. 10 is not nearly as smooth as that of the proposed approach and, consequently, yields a significantly higher median error of 12.5 degrees. These results indicate the benefit of the more integrated spatiotemporal processing approaches **STE-local** and **STQ-local** as well as the more reliable spatiotemporal orientation measurement of **STE** in comparison to standard first-order derivatives underlying the flow estimates of **noST-LK**.

Figure 11 shows recovered 3D scene flow for the *Lab 2* sequence. While this sequence does not have associated scene flow groundtruth, it illustrates results in a more complicated situation than *Lab3*. It also has more complicated flow than *Lab1*, as it involves both translation and rotation. The flow vector confidence measure, ζ , (19) also is displayed. The recovered motion is qualitatively correct and appears quite smooth considering that no explicit optimization over flow vectors has been attempted. The 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; further, the background is characterized with very light pink in the d -motion map signalling the camera moving forward. It also is seen that the confidence measure, ζ , reports reasonable values, e.g., highest in areas with enough image texture to yield adequate variation within spatiotemporal oriented

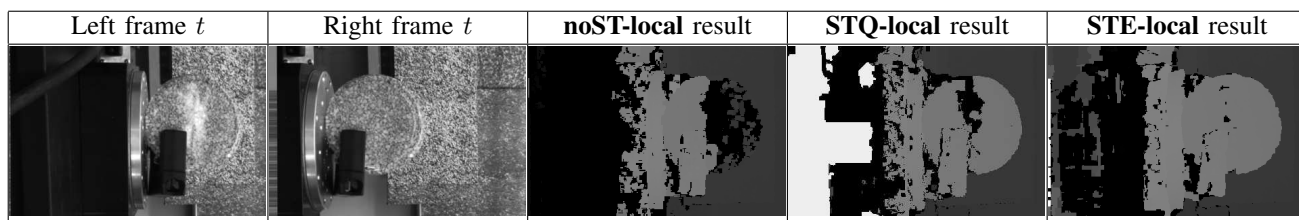


Fig. 8. Input frames and disparity maps for *Lustre* dataset.

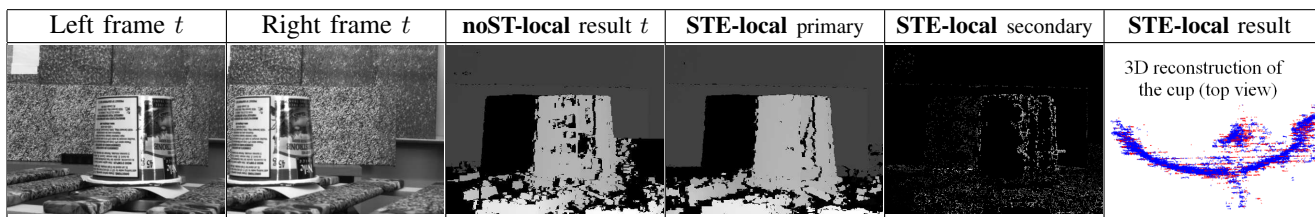


Fig. 9. Example input frames and recovered disparity for *Bino-spec* data set.

energy distribution, (2), to support motion estimates and low in untextured regions, such as the black backgrounds.

Figure 11 also shows results for **noST-LK**. In general, the resulting 3D flow maps from **STE-local** and **noST-LK** exhibit similar qualitative patterns; however, **STE-local** yields consistently smoother maps with fewer gross errors (especially in the depth component). Finally, the results for the **STQ-local** approach to scene flow estimation on *Lab2* were presented previously [49]. While not replicated here due to space, visual comparison shows that they are qualitatively similar to those of **STE-local** shown in Fig. 11 and thereby consistent with the quantitative comparison presented in conjunction with Fig. 10.

4 DISCUSSION

This paper has described a novel approach to spacetime stereo and motion recovery based on spatiotemporal oriented energy distributions as match primitives. Several important contributions can be explicitly recounted. First, since the primitives and match cost inherently involve the temporal dimension, the resulting disparity estimates naturally exhibit temporal coherence. Second, matches that are ambiguous when considering only spatial pattern are resolved through the inclusion of temporal information. Third, a unique approach to multilayer disparity estimation (essential for robust processing of (semi)transparent and specularly reflecting surfaces) is developed based on allowing subsets of the orientation measurements to support different disparity estimates. Fourth, a method for direct recovery of 3D disparity flow from matched spatiotemporal oriented energy distributions is developed. Fifth, prototype recovery algorithms have been designed and implemented in real-time on commodity GPUs. In comparison to alternative approaches, these benefits have been documented qualitatively and quantitatively on both publicly available and novel data sets. Video results for all datasets are presented in supplementary material downloadable from <http://www.cse.yorku.ca/~sizints/TPAMI-STE-2013.mp4>

The present work can be perceived as a logical continuation to previous work using spatiotemporal orien-

tation as encapsulated in the spatiotemporal quadric element (stequel) [49]. From a theoretical point of view the present approach makes more complete use of available spatiotemporal orientation information, as it does not collapse (potentially multimodal) orientation distributions into a quadric approximation. Importantly, this theoretical advantage has been shown to have practical ramifications, especially in the resolution of disparity in the vicinity of surface discontinuities and the explicit recovery of multilayer estimates when appropriate (e.g., transparency and specular reflection). Similarly, the proposed 3D flow estimation also has potential to support recovery of multiple flow vectors by considering multiple local minima in its optimization function. More generally, it appears that the proposed approach is the only research on spacetime stereo to consider multilayer disparity estimation.

ACKNOWLEDGMENTS

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

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] J. R. Bergen, P. J. Burt, R. Hingorani, and S. Peleg. A three-frame algorithm for estimating two-component image motion. *TPAMI*, 14(9):886–896, 1992.
- [4] J. Bigun. *Vision with Direction*. Springer, 1998.
- [5] M. J. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *CVIU*, 61(1):75–104, 1996.
- [6] M. Bleyer and M. Gelautz. Temporally consistent disparity maps from uncalibrated stereo videos. In *ISPA*, pages 383–387, 2009.
- [7] M. Borga and H. Knutsson. Estimating multiple depths in semi-transparent stereo images. In *SCIA*, 1999.
- [8] M. Z. Brown, D. Burschka, and G. D. Hager. Advances in computational stereo. *TPAMI*, 25(8):993–1008, 2003.
- [9] K. J. Cannons and R. P. Wildes. The applicability of spatiotemporal oriented energy features to region tracking. *TPAMI*, 36(4):784–796, 2014.
- [10] O. Chomat, J. Martin, and J. Crowley. A probabilistic sensor for the perception and the recognition of activities. In *ECCV*, 2000.
- [11] J. Davis, R. Ramamoorthi, and S. Rusinkiewicz. Spacetime stereo: A unifying framework for depth from triangulation. *TPAMI*, 27(2):296–302, 2005.

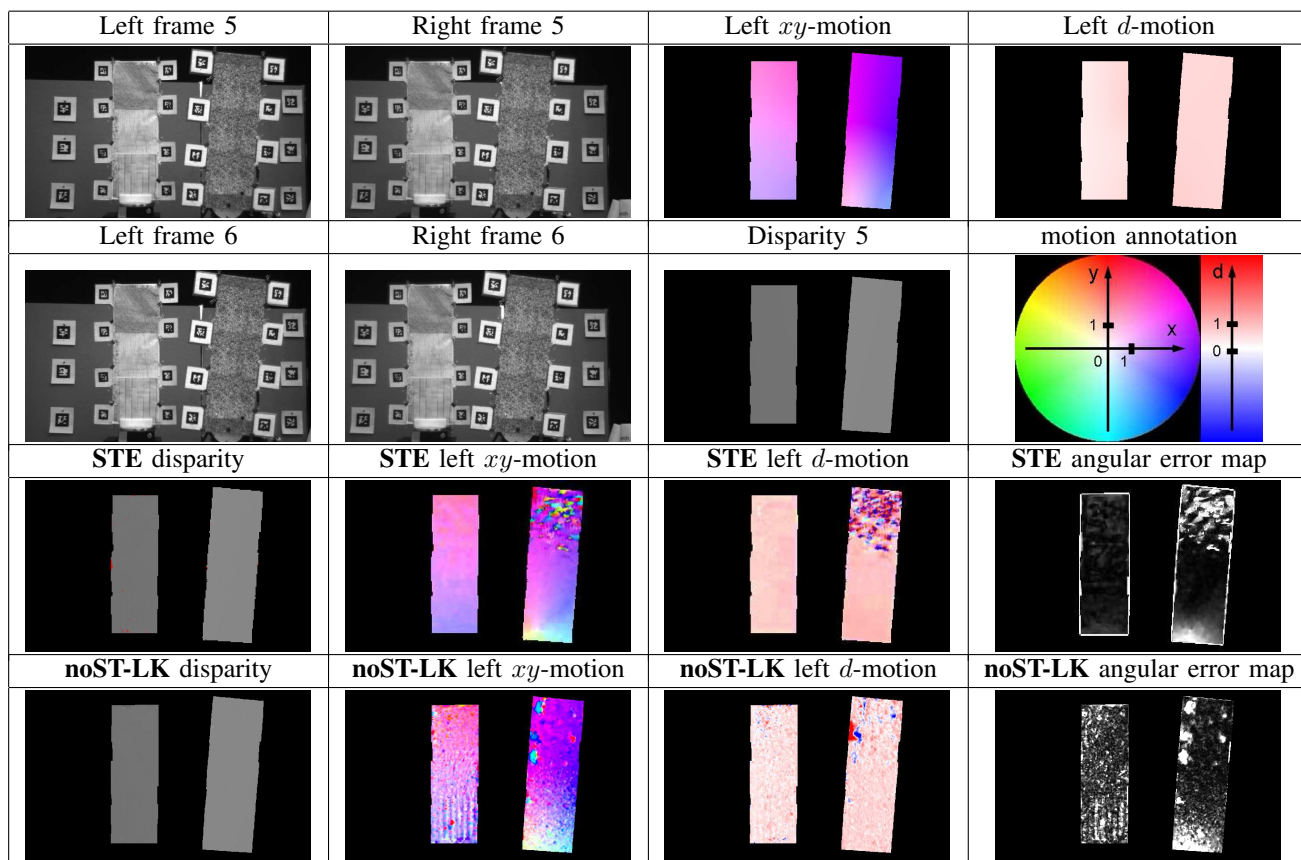


Fig. 10. *Lab 3* dataset and motion estimation results. Top two rows: Left half shows the original intensity images for time consecutive frames, while right half shows disparity and colour-coded flow components and the associated annotation chart. Bottom two rows: Motion and disparity estimation results recovered at the middle frame using STE-local and noST-LK. Angular error is plotted with black through white depicting 0 to 90 or more degrees.

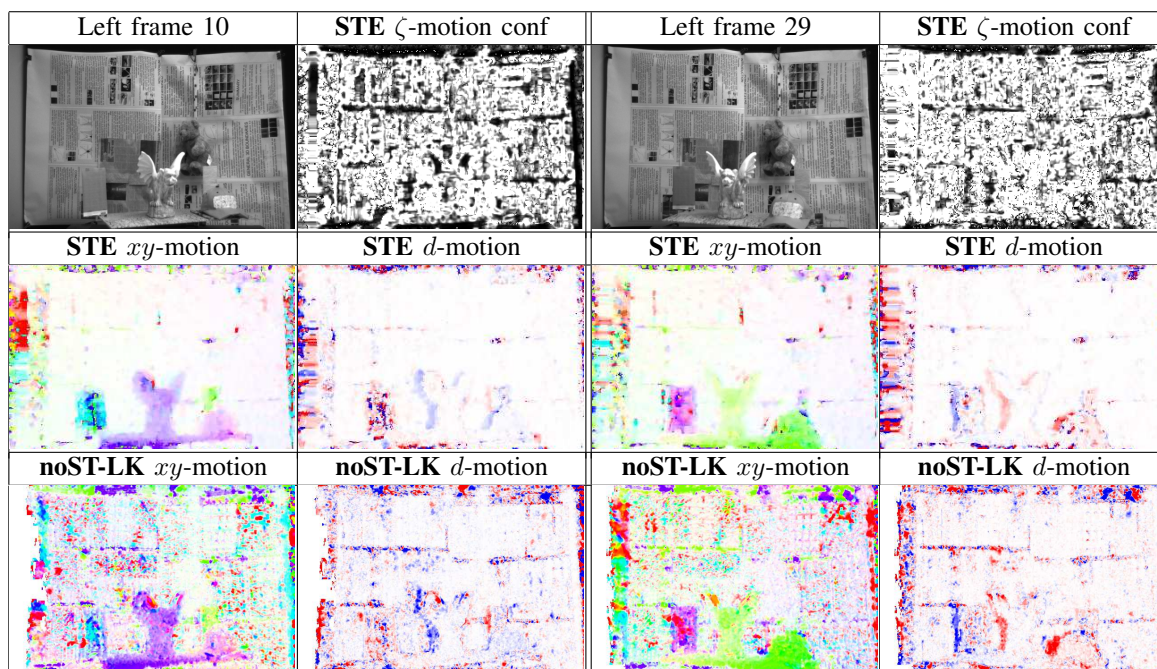


Fig. 11. Example Motion Estimation Results for *Lab2* at Different Time Frames. Top row: raw frame and motion confidence ζ for STE-local (brighter values corresponds to increased confidence). Middle row: xy and d components of 3D motion estimation for STE-local. Bottom row: xy and d components of 3D motion estimation for noST-LK. 3D flow vector colour coding is consistent with that originally presented in Fig. 10.

- [12] D. Demirdjian and T. Darrell. Using multiple-hypothesis disparity maps and image velocity for 3D motion estimation. *IJCV*, 47:219–228, 2002.
- [13] K. G. Derpanis, M. Sizintsev, K. Cannons, and R. P. Wildes. Efficient action spotting and recognition based on a spacetime oriented structure representation. *TPAMI*, 35(3):527–540, 2013.
- [14] K. G. Derpanis and R. P. Wildes. Early spatiotemporal grouping with a distributed oriented energy representation. In *CVPR*, 2009.
- [15] K. G. Derpanis and R. P. Wildes. The structure of multiclaicite motions in natural imagery. *TPAMI*, 32(7):1310–1316, 2010.
- [16] K. G. Derpanis and R. P. Wildes. Spacetime texture representation and recognition based on a spatiotemporal orientation analysis. *TPAMI*, 34(6):1193–1205, 2011.
- [17] R. O. Duda and P. E. Hart. Use of the hough transformation to detect lines and curves in pictures. *CACM*, 15:11–15, 1972.
- [18] 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.
- [19] W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *TPAMI*, 13(9):891–906, 1991.
- [20] M. Gong. Enforcing temporal consistency in real-time stereo estimation. In *ECCV*, pages 564–577, 2006.
- [21] G. Granlund and H. Knutsson. *Signal Processing for Computer Vision*. Kluwer, 1995.
- [22] K. J. Hanna and N. E. Okamoto. Combining stereo and motion analysis for direct estimation of scene structure. In *ICCV*, pages 357–365, 1993.
- [23] B. K. P. Horn. *Robot Vision*. The MIT Press, 1986.
- [24] P. V. C. Hough. Machine analysis of bubble chamber pictures. In *Proc. Int. Conf. High Energy Accel. and Instr.*, 1959.
- [25] F. Huguet and F. Devernay. A variational method for scene flow estimation from stereo sequences. In *ICCV*, pages 1–7, 2007.
- [26] C. H. Hung, L. Xu, and J. Jia. Consistent binocular depth and scene flow with chained temporal profiles. *IJCV*, 102(1-3):271–292, 2013.
- [27] M. Isard and J. MacCormick. Dense motion and disparity estimation via loopy belief propagation. In *ACCV*, pages 32–41, 2006.
- [28] H. Jiang, H. Liu, P. Tan, G. Zhang, and H. Bao. 3d reconstruction of dynamic scenes with multiple handheld cameras. In *ECCV*, 2012.
- [29] 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.
- [30] V. Kolmogorov and R. Zabih. Computing visual correspondence with occlusions using graph cuts. In *ICCV*, pages 508–515, 2001.
- [31] E. S. Larsen, P. Mordohai, M. Pollefeys, and H. Fuchs. Temporally consistent reconstruction from multiple video streams using enhanced belief propagation. In *ICCV*, pages 1–8, 2007.
- [32] S.-B. Lee and Y.-S. Ho. Temporally consistent depth map estimation using motion estimation for 3DTV. In *WAIT*, pages 149–154, 2010.
- [33] C. Lei, X. D. Chen, and Y. Yang. A new multiview spacetime-consistent depth recovery framework for free viewpoint video rendering. In *ICCV*, pages 1570–1577, 2009.
- [34] 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.
- [35] S. Malassiotis and M. G. Strintzis. Model-based joint motion and structure estimation from stereo images. *CVIU*, 65(1):79–94, 1997.
- [36] 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.
- [37] Middlebury College Stereo Vision Page. <http://www.middlebury.edu/stereo/>, 2013.
- [38] J. Neumann and Y. Aloimonos. Spatio-temporal stereo using multi-resolution subdivision surfaces. *IJCV*, 47(1-3):181–193, 2002.
- [39] OpenCL by Khronos Group. www.khronos.org/opencl.
- [40] P. Pearce and S. Pearce. *Polyhedra primer*. Van Nostrand, 1979.
- [41] J.-P. Pons, R. Keriven, O. D. Faugeras, and G. Hermonillo. Variational stereovision and 3D scene flow estimation with statistical similarity measures. In *ICCV*, pages 597–602, 2003.
- [42] C. Richardt, D. Orr, I. Davies, A. Criminisi, and N. A. Dodgson. Real-time spatiotemporal stereo matching using the dual-cross-bilateral grid. In *ECCV*, 2010.
- [43] H. Scharr and R. Kusters. A linear model for simultaneous estimation of 3D motion and depth. In *WVM*, pages 220–225, 2002.
- [44] H. Scharr and T. Schuchert. Simultaneous motion, depth and slope estimation with a camera-grid. In *WVM*, pages 81–88, 2006.
- [45] D. Scharstein and R. Szeliski. Taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47:7–42, 2002.
- [46] M. Shizawa. Direct estimation of multiple disparities for transparent multiple surfaces in binocular stereo. In *ICCV*, pages 447–454, 1993.
- [47] M. Sizintsev and R. P. Wildes. Coarse-to-fine stereo vision with accurate 3D boundaries. *IVC*, 28(3):352–366, 2010.
- [48] M. Sizintsev and R. P. Wildes. Spatiotemporal oriented energies for spacetime stereo. In *ICCV*, 2011.
- [49] M. Sizintsev and R. P. Wildes. Spatiotemporal stereo and motion via stequel matching. *TPAMI*, 34(6):1206–1219, 2012.
- [50] G. Stein and A. Shashua. Direct estimation of motion and scene structure from a moving stereo rig. In *CVPR*, pages 211–218, 1998.
- [51] G. Strang. *Linear Algebra and its Applications*. HBJ, 1988.
- [52] C. Strecha and L. van Gool. Motion-stereo integration for depth estimation. In *ECCV*, pages 170–185, 2002.
- [53] 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.
- [54] Y. Tsin, S. B. Kang, and R. Szeliski. Stereo matching with linear superposition of layers. *TPAMI*, 28(2):290–301, 2006.
- [55] L. Valgaerts, A. Bruhn, H. Zimmer, J. Weickert, C. Stoll, and C. Theobalt. Joint estimation of motion, structure and geometry from stereo sequences. In *ECCV*, 2010.
- [56] A. Wedel, C. Rabe, T. Vaudrey, T. Brox, U. Franke, and D. Cremers. Efficient dense scene flow from sparse or dense stereo data. In *ECCV*, volume 1, pages 739–751, 2008.
- [57] O. Williams, M. Isard, and J. MacCormick. Estimating disparity and occlusions in stereo video. In *CVPR*, pages 250–257, 2005.
- [58] W. Xiong and J. Jia. Stereo matching on objects with factional boundary. In *CVPR*, 2007.
- [59] M. Yang, X. Cao, and Q. Dai. Multiview video depth estimation with spatial-temporal consistency. In *BMVC*, 2010.
- [60] W. Yang, G. Zhang, H. Bao, J. Kim, and H. Lee. Consistent depth maps recovery from a trinocular video sequence. In *CVPR*, 2012.
- [61] A. Zaharescu and R. P. Wildes. Anomalous behaviour detection using spatiotemporal oriented energies, subset inclusion histogram comparison and event-driven processing. In *ECCV*, 2010.
- [62] G. Zhang, J. Jia, T. Wong, and H. Bao. Consistent depth map recovery from a video sequence. *TPAMI*, 31(6):974–988, 2009.
- [63] L. Zhang, B. Curless, and S. M. Seitz. Spacetime stereo: Shape recovery for dynamic scenes. In *CVPR*, pages 367–374, 2003.
- [64] Y. Zhang and C. Kambhamettu. On 3D scene flow and structure estimation. In *CVPR*, volume 2, pages 778–785, 2001.
- [65] J. Zhu, M. Liao, R. Yang, and Z. Pan. Joint depth and alpha matter optimization via fusion of stereo and time-of-flight sensor. In *CVPR*, pages 453–460, 2009.



Mikhail Sizintsev (Member, IEEE) received the BSc (Honours), MSc and PhD degrees in computer science from York University, Toronto, Canada in 2004, 2006 and 2012, respectively. He spent the summer 2009 at Sarnoff Corporation in Princeton, New Jersey as an intern developing GPU-based stereo systems for augmented reality applications. Currently, he is a computer scientist at SRI International in Princeton, New Jersey. His major areas of research include stereo, motion, augmented reality and multi-sensory navigation.



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 Lab. 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.