Spatiotemporal Oriented Energies for Spacetime Stereo

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

Abstract

This paper presents a novel approach to recovering temporally coherent estimates of 3D structure 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. The approach has been implemented with real-time performance on commodity GPUs. Empirical evaluation shows that the approach yields qualitatively and quantitatively superior disparity estimates in comparison to various alternative approaches, including the ability to provide accurate multilayer estimates in the presence of (semi)transparent and specular surfaces.

1. Introduction

Binocular stereo is one of the fundamental and most widely researched topics in computer vision. Broadly stated, given a pair of spatially separated 2D projections of a scene, the goal is to recover the unknown third dimension of distance between the sensor and scene. Significantly, many applications allow for image acquisition over time and thereby allow for incorporation of the temporal dimension into processing. Ideally, recovered 3D structure should be temporally consistent and respect scene dynamics. Furthermore, the addition of temporal information has the potential to resolve stereo matches that might be ambiguous when only instantaneous binocular views are considered.

In response to these observations, the present paper proposes a novel approach to spacetime stereo that centers on the idea of representing video in terms of a distribution of 3D orientation measurements in visual spacetime, (x, y, t). The measurements are recovered via application of a bank of orientation tuned spatiotemporal filters separately to the left and right image streams for subsequent matching. By capturing both spatial and temporal structure in this unified fashion, the cues combine to drive matching for resolution of situations that might be ambiguous when either source is considered alone and to increase temporal continuity. Further, by allowing subsets of the orientation measurements to support different disparity estimates, a natural approach to recovering multilayer disparity estimates arises that yields accurate recovery in the presence of depth discontinuities, (semi)transparency and specular reflections.

Various attempts have been made to understand how the availability of temporal information can enhance binocular stereo processing. Considerable work has addressed simultaneous structure and motion estimation that combines intraframe (spatial) and interframe (temporal) image pairwise constraints in various fashions (*e.g*. [27, 21, 14, 29]). However, the following review concentrates on research that uses temporal information primarily to enhance disparity estimation (rather than 3D motion processing), as such work is most closely related to the current research. 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 [18]. Other approaches reinforce disparity hypotheses by propagating correlation scores from the previous frame using optical flow [13]. Still other approaches consider temporal information by extending a regularizing spatial MRF grid to include time and thereby allow for smoothing along the temporal direction, variously respecting flow displacements [17] and either accounting for change detection [31] or not [19].

The proposed approach explicitly combines spatial and temporal (*i.e*. spatiotemporal) support in stereo matching; thus, previous research with similar considerations is of particular interest. One method was initially developed in conjunction with temporally varying structured lighting [9]. Other work generalized this approach to model temporal disparity change [32]. Still other work extends the notion of spatially adaptive aggregation to include the temporal dimension [22]. Most closely related to the proposed ap-

proach is previous work that used measurements of spatiotemporal orientation as the basis for stereo matching [24]. That work encapsulated spacetime orientation in the spatiotemporal quadric or stequel (also referred to as the orientation tensor and covariance matrix [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.

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 [11], tracking [7] and activity recognition [8], it appears that no previous work has applied this approach directly to spacetime stereo. Previous work has made use of purely spatial orientation measurements in stereo matching [15], but did not consider the temporal dimension.

In the light of previous work, the outstanding contributions of the proposed approach are as follows. First, a novel approach to spacetime disparity estimation is proposed based on direct matching of a distribution of image spacetime orientation measurements. In distinction from an alternative spacetime stereo method [24], the present approach eschews collapsing orientation measurements into a quadric approximation and thereby makes fuller use of available information. Second, the first approach to recovering multilayer disparity estimates from spacetime 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.* [23, 28]) 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 [10]. Third, the approach is realized in local and global stereo matchers with real-time GPU-based performance for the local version. Fourth, the developed implementations have been subject to extensive qualitative and quantitative empirical evaluation.

2. Technical Approach

2.1. Background

2.1.1 Spatiotemporal orientation correspondence

Local oriented measurements in image spacetime have visual significance and thereby are an appropriate primitive for spacetime stereo. 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).

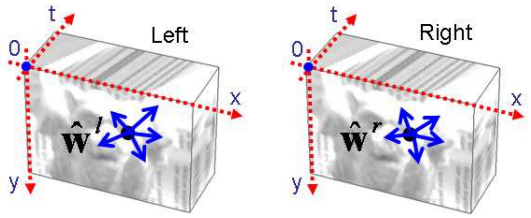

Figure 1. The spatiotemporal orientation correspondence constraint, (1), describes the relationship between arbitrary orientations in correspondence, $\hat{\mathbf{w}}^l$ and $\hat{\mathbf{w}}^r$, subject to binocular viewing of a slanted surface undergoing arbitrary motion in the world relative to the cameras. Depicted are several different orientation directions that might be considered across views.

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 originally elsewhere [24] and is summarized here to provide necessary groundwork. Let unit vectors $\hat{\mathbf{w}}^l$ and $\hat{\mathbf{w}}^r$ (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. 1. These orientations are related as

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

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. In the following, the basic relationship between binocularly corresponding image spacetime orientations, (1), will be referred to as the *spatiotemporal orientation correspondence constraint*.

2.1.2 Measuring local spatiotemporal orientation

To exploit the spatiotemporal orientation correspondence constraint, (1), one must commit to a particular approach to making 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 [12], 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

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

with $*$ denoting convolution.

Significantly, most practical uses of energy filtering of the form (2) 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 is realized via pointwise division by the sum of the N local energy measurements at a point:

$$\hat{E}(\mathbf{x};\hat{\mathbf{w}}_i) = E(\mathbf{x};\hat{\mathbf{w}}_i) / \left(\sum_{j=1}^{N} E(\mathbf{x};\hat{\mathbf{w}}_j) \right). \quad (3)$$

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 Hilbert transforms are used; so, $N = 10$ is required [12], with their orientations chosen to uniformly sample 3D orientation as the normals to the faces of an icosahedron with antipodal directions identified [4].

2.2. Binocular spatiotemporal orientation error

With both the relationship between binocularly corresponding spatiotemporal orientations, (1), and a method for measuring local orientations, (3), 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, (1), 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, (3), 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 (3) as

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

which by (1) evaluates to

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

The error function, (5), is minimized by setting the corresponding gradient with respect to $\mathbf{h} = [h_1\ h_2\ h_3]^\top$ to zero and subsequently solving for $\mathbf{h}$. Each error component $\mathcal{E}_i^2$ 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 $\mathbf{h}_0 = [0,0,0]$ to arrive at the simpler form

$$\tilde{\mathcal{E}}_i(\mathbf{x}^l,\mathbf{x}^r) = \mathcal{E}_i(\mathbf{x}^l,\mathbf{x}^r;\mathbf{h}_0) + \nabla\mathcal{E}_i^\top(\mathbf{x}^l,\mathbf{x}^r;\mathbf{h}_0)\mathbf{h} \quad (6)$$

for the error associated with each orientation $\mathbf{w}_i$.

Algorithm 1 Spacetime multilayer disparity estimation

```
Let D be the set of disparities considered
initialize voting array V corresponding to D to zero
for all directions ŵ_i do
  for all disparities d ∈ D do
    compute E_i^2 using (12)
  end for
  let d_i = arg min_d E_i
  update V(d_i) = V(d_i) + 1
end for
for all disparity hypotheses d from D do
  if V(d) ≥ voting threshold λ_v then
    Declare d as one of the disparities
  end if
end for
```

2.3. Spatiotemporal orientation match cost

In this section, two methods are presented for assigning a cost to matching points $\mathbf{x}^l = (x^l, y^l, t)^\top$ and $\mathbf{x}^r = (x^l + d, y^l, t)^\top$ across a binocular view according to disparity estimate, d.

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

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

where 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

$$B_{i,m} = \frac{\partial \mathcal{E}_i(\mathbf{x}^l,\mathbf{x}^r;\mathbf{h}_0)}{\partial h_m}, \quad \text{and} \quad b_i = \mathcal{E}_i(\mathbf{x}^l,\mathbf{x}^r;\mathbf{h}_0), \quad (8)$$

so that each row of B captures the error contribution of a particularly sampled direction $\hat{\mathbf{w}}_i$. A solution for $\mathbf{h}$ can be obtained by following standard linear algebraic manipulations [26] as $\mathbf{h} = (\mathsf{B}^\top\mathsf{B})^{-1}\mathsf{B}^\top\mathbf{b}$ with residual error

$$\tilde{\mathcal{E}}^2 = \sum_{i=1}^{N} \tilde{\mathcal{E}}_i^2 = \left(\mathsf{B}\left(\mathsf{B}^\top\mathsf{B}\right)^{-1}\mathsf{B}^\top\mathbf{b} - \mathbf{b} \right)^2. \quad (9)$$

Thus, for any given disparity, d, the cost associated with matching $\mathbf{x}^l$ with $\mathbf{x}^r$ is taken as the residual, (9).

It is not necessary to combine all orientation measurements into a single system of equations in support of a single minimal cost disparity estimate. Alternatively, multiple disparity estimates can be recovered at a point or over a spatial region by allowing subsets of measurements to contribute to different estimates, as might be useful in cases of (semi)transparencies, reflections and even near surface discontinuities. In particular, the second method for assigning costs to disparities operates by aggregating orientation measurements over a spatial support window so that each orientation, $\hat{\mathbf{w}}_i$, can form its own (overdetermined) least-squares system to define the error cost $\mathcal{E}_i(\mathbf{x}^l,\mathbf{x}^r;\mathbf{h}_0)$ for a disparity estimate. In this formulation, consideration is of the system

$$\mathsf{C}\mathbf{h} = \mathbf{c}, \quad (10)$$

where C is a $Q \times 3$ matrix, $\mathbf{c}$ is an $Q \times 1$ vector and Q is the number of points in the spatial aggregation region, with

$$C_{q,m} = \frac{\partial \mathcal{E}_i(\mathbf{x}_q^l, \mathbf{x}_q^r; \mathbf{h}_0)}{\partial h_m}, \text{ and } c_q = \mathcal{E}_i(\mathbf{x}_q^l, \mathbf{x}_q^r; \mathbf{h}_0), \quad (11)$$

so that each row of C captures the error contribution of particular points $\mathbf{x}_q^l$ and $\mathbf{x}_q^r$ that ranges over a spatial aggregation region of Q total points. Similarly to (7), the solution is $\mathbf{h} = (\mathsf{C}^\top \mathsf{C})^{-1} \mathsf{C}^\top \mathbf{c}$ with residual error

$$\tilde{\mathcal{E}}_i^2 = \left(\mathsf{C} \left(\mathsf{C}^\top \mathsf{C} \right)^{-1} \mathsf{C}^\top \mathbf{c} - \mathbf{c} \right)^2. \quad (12)$$

In this case, each measured orientation, $\hat{\mathbf{w}}_i$, can provide its own cost for any given disparity, d, in terms of the residual (12). In turn, each measured orientation can "vote" for its own lowest cost disparity and all disparities that receive greater than a threshold number of votes can be considered valid and thereby allow for multiple layer disparity estimation. This approach is encapsulated in Alg. 1.

The first method for assigning cost to disparity estimates, (9), 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, [24]; however, the proposed approach makes more complete use of available orientation information by eschewing its collapse to the quadric. The second method, (12), in conjunction with Alg. 1, allows for multilayer disparity estimates over a window of match aggregation by more fully exploiting the availability of multiple orientation measurements.

2.4. Temporal flicker cue

Temporal continuity in the captured binocular imagery can break down (*e.g.* motions so large that they alias, flashing lights that yield instantaneous brightness change). In such situations, it is useful to restrict orientation-based matching to consider only orientation filtering in space (i.e., neglect support in the temporal domain) and thereby avoid contamination from confusing temporal information.

Along these lines, an interesting aspect of spatiotemporal oriented energy measurements is that individual directions can be associated with qualitative descriptors of image dynamics. In particular, filters oriented in directions orthogonal to t-axis (*e.g.*, $[1\ 0\ 0]^\top$) are matched to temporal change and their response is indicative of infinite/very fast velocity and instantaneous intensity change, *i.e.* breakdown in temporal continuity. Normalized energy response along the spanning set of these directions is referred to as *flicker* [30, 11], $\mathcal{F}$, and can be readily computed as a linear combination of the basis filter responses, (3), implying that intensity isocontours of minimal brightness variation must lie in the xy plane:

$$\mathcal{F}(\mathbf{x}) = 1 - \left[\hat{E}(\mathbf{x}; [1\ 0\ 0]^\top) + \hat{E}(\mathbf{x}; [0\ 1\ 0]^\top) \right]. \quad (13)$$

Thus, $0 \leq \mathcal{F} \leq 1$, with $\mathcal{F} \to 1$ as the signal becomes pure temporal change. In the following, local measurements of $\mathcal{F}$ that exceed a threshold are used to detect locations where 3D spatiotemporal, (x, y, t), orientation matching is switched to purely 2D spatial, (x, y), orientation matching. Importantly, there is no need to recompute 2D filter responses; they can be extracted from the already computed 3D filtering operations by reusing the intermediate results of separable 1D convolutions along the x and y axes [12].

3. Experimental evaluation

The proposed approach has been realized in software implementations that input synchronized and rectified binocular videos, I^l, I^r, 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 (9) 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 [25]. The global algorithm is a graph-cuts matcher [16]. These particular matchers were chosen because they have been used previously in realizing the stequel approach to spacetime stereo [24] and thereby allow for direct comparison.

The local method makes use of the per orientation match cost (12) with Alg. 1 to support recovery of multiple layer disparity estimates. In all cases spatial aggregation is 5×5 and voting threshold $\lambda_v = 4$. The global method makes use of the across orientations match cost, (9), 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, (9), also was embedded in the local method; it was found that results were comparable to those shown here for single layer disparity estimates and are not given explicitly for the sake of space. For both implementations, subpixel estimation was performed as post-processing using a Lucas-Kanade type refinement [1] specialized to the proposed spatiotemporal match costs, (9) and (12), as done analogously in previous comparable work [24].

The local algorithm, **STE-local**, is well suited to parallel computation and therefore has been implemented in OpenCL [20] 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. 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 cor-

relation match cost, as realized in both local adaptive, coarse-to-fine block [25] and global graph-cut [16] algorithms; these methods will be denoted **noST-local** and **noST-global**, resp. Second, comparison is made to the most closely related stequel-based matching, again with both local and global instantiations [24], 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 [32]. As with all others instances, this approach has been implemented within the same local [25] and global [16] matchers, denoted **Zhang-local** and **Zhang-global**, resp.

Six binocular video data sets are used as input. The first two are the *Lab 1* and *Lab 2* videos originally presented elsewhere [24]. 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 Fig. 2; see supplemental material for video results. Disparity maps are shown only for **STE** and **Zhang**, as recovered disparity maps for the other approaches are available elsewhere [24] and are suppressed here in the interest of space; error statistics are shown for all approaches. 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*. in the piecewise planar regions of *Lab 1*) 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, as evidenced in the error statistics near discontinuities. 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 *Lab 1*, **Zhang** does help disambiguate matches in the camouflage (lower left) and epipolar-aligned texture regions relative to **noST**. Its performance on *Lab 2* 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 [32]. 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.

An important distinguishing point of Alg. 1 in comparison to all previous spatiotemporal stereo algorithms is the ability to deal with multilayer disparities at a point. The third test data set, *Transparency*, illustrates the case of semi-transparency. This 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. 3. 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 Alg. 1 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 Alg. 1 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

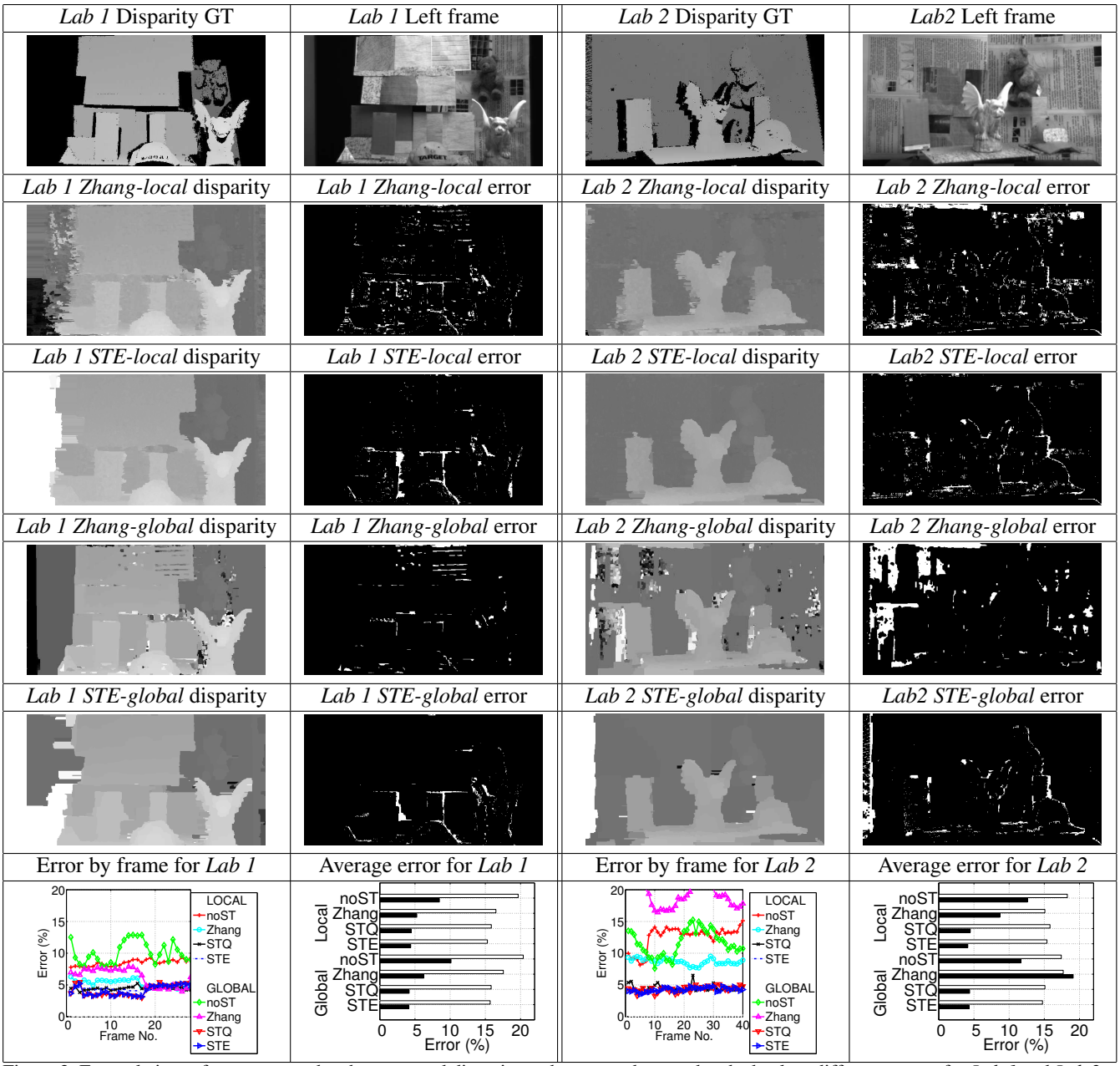

Figure 2. Example input frames, groundtruth, recovered disparity and recovered-groundtruth absolute difference error for *Lab 1* and *Lab 2*. For summary statistics, an error is taken as greater than 1 pixel discrepancy between recovered and groundtruth disparity. Bar plots show average error across 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.

also was applied to the *Transparency* data set. This matcher makes use of robust, parametric layer estimation [5] and was applied to the same left and right frames used to present results for the proposed approach, **STE-local**. The results are plotted as green planes in the perspective plot of Fig. 3. 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.

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 fourth data set, *Lustre*, deals with the case of "binocular lustre" where a specular reflection is present in one of the two views and totally absent in

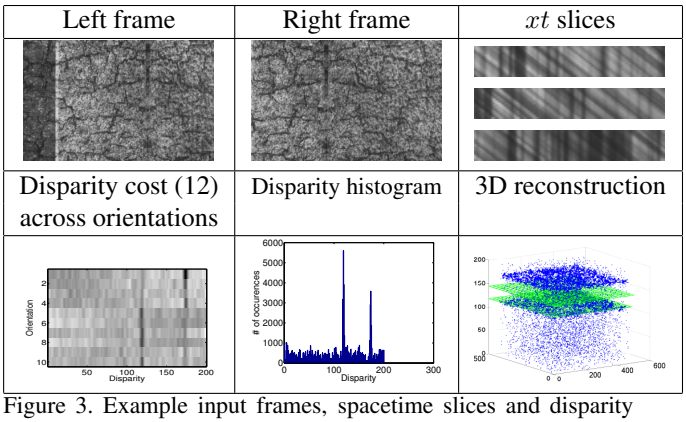

Figure 3. Example input frames, spacetime slices and disparity estimation results for *Transparency*. See text for details.

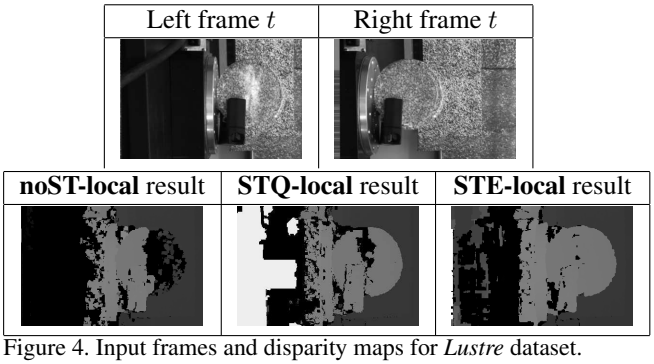

Figure 4. Input frames and disparity maps for *Lustre* dataset.

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. 4. Spatiotemporal stequel matcher **STQ-local** is able to reasonably capture the surfaces 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.

The fifth data set, *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. 5. 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

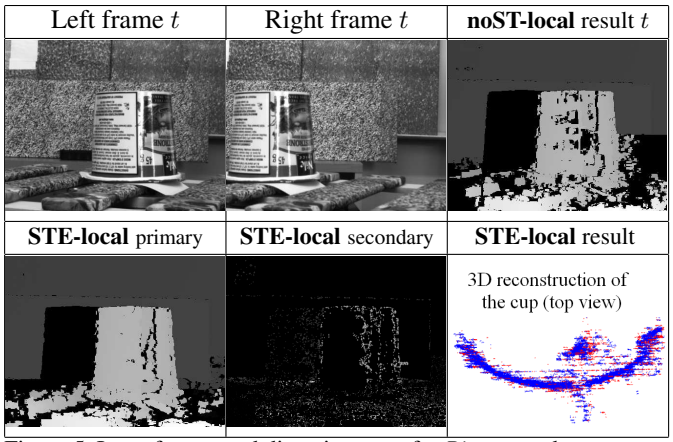

Figure 5. Input frames and disparity maps for *Bino-spec* dataset.

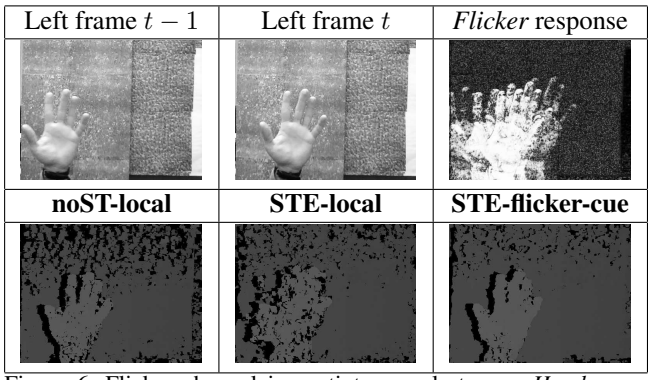

Figure 6. Flicker channel in spatiotemporal stereo. *Handwave* dataset: Example time-consecutive frames, flicker response and disparity maps for various methods.

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_v, the "secondary estimate" map shows other disparities whose number of votes also surpassed λ_v, 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 the mirror reflection with respect to a 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.

As already demonstrated, spacetime stereo offers a set of advantages over spatial-only stereo. Meanwhile, very large motions that result in temporal aliasing (*e.g.* situations when the displacement of the object is larger than the size of the object itself) creates significant difficulty for spacetime methods, since temporal continuity breaks down. Dataset *Handwave* (Fig. 6) shows a simple scenario of a rapidly moving hand and disparity maps processed with **noST-local**

and **STE-local**. Here, the **STE** algorithm behaves worse than traditional stereo **noST**. Fortunately, places of excessively rapid motion can be detected using flicker, $\mathcal{F}$, and matching can be restricted to spatial only filtering at such points, as described in Sec. 2.4. The results in Fig. 6 labeled **STE-flicker-cue** were generated by switching from spatiotemporal to purely spatial matching when $\mathcal{F} > 0.5$.

4. Discussion

This paper has described a novel approach to spacetime stereo using spatiotemporal oriented energy distributions as match primitives. Points of distinction include the following: (i) resulting disparity estimates naturally exhibit temporal coherence, as the primitives and match cost inherently involve the temporal dimension; (ii) matches that are ambiguous when considering only spatial pattern are resolved through the inclusion of temporal information; (iii) by allowing subsets of the orientation measurements to support different disparity estimates, an approach to multilayer disparity estimation is realized, *e.g.* as useful in the presence of (semi)transparent and specularly reflecting surfaces; (iv) a method for detecting and treating points where temporal continuity breaks down (flicker) is presented; (v) the approach is amenable to real-time implementation on commodity GPUs. In comparison to alternative approaches, these benefits have been documented qualitatively and quantitatively on both publicly available and novel data sets.

Perhaps the most closely related approach is previous work using spatiotemporal orientation as encapsulated in the spatiotemporal quadric element (stequel) [24]. From a theoretical point of view the proposed approach makes more complete use of available spatiotemporal orientation information, as it does not collapse (potentially multimodal) orientation distributions into a quadric approximation. 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). More generally, it appears that the proposed approach is the only research on spacetime stereo to consider multilayer disparity estimation.

Acknowledgements

This research was supported in part by NSERC and MDA Space Missions.

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] K. Cannons and R. Wildes. Visual tracking using pixelwise spatiotemporal oriented energy representation. In *ECCV*, 2010.

[8] O. Chomat, J. Martin, and J. Crowley. A probabilistic sensor for the perception and the recognition of activities. In *ECCV*, 2000.

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

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

[11] K. Derpanis and R. Wildes. Dynamic texture recognition based on distributions of spacetime oriented structure. In *CVPR*, 2010.

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

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

[14] M. Isard and J. MacCormick. Dense motion and disparity estimation via loopy belief propagation. *ACCV*, pages 32–41, 2006.

[15] 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.

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

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

[18] S.-B. Lee and Y.-S. Ho. Temporally consistent depth map estimation using motion estimation for 3DTV. In *WAIT*, pages 149–154, 2010.

[19] 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.

[20] OpenCL by Khronos Group. www.khronos.org/opencl.

[21] 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.

[22] 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.

[23] M. Shizawa. Direct estimation of multiple disparities for transparent multiple surfaces in binocular stereo. *ICCV*, 1:447–454, 1993.

[24] M. Sizintsev and R. Wildes. Spatiotemporal stereo via spatiotemporal quadric element (stequel) matching. *CVPR*, 2009.

[25] M. Sizintsev and R. P. Wildes. Coarse-to-fine stereo vision with accurate 3D boundaries. *IVC*, 28(3):352–366, 2010.

[26] G. Strang. *Linear Algebra and its Applications*. HBJ, 1988.

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

[28] Y. Tsin, S. B. Kang, and R. Szeliski. Stereo matching with linear superposition of layers. *TPAMI*, 28(2):290–301, 2006.

[29] 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.

[30] R. Wildes and J. Bergen. Qualitative spatiotemporal analysis using an oriented energy representation. *ECCV*, 1:786–784, 2000.

[31] O. Williams, M. Isard, and J. MacCormick. Estimating disparity and occlusions in stereo video sequences. *CVPR*, 1:250–257, 2005.

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