



Spatiotemporal Stereo via Spatiotemporal Quadric Element (Stequel) Matching

Mikhail Sizintsev

Richard P. Wildes

Technical Report CSE-2008-04

July 11, 2008

Department of Computer Science and Engineering
4700 Keele Street Toronto, Ontario M3J 1P3 Canada

Spatiotemporal Stereo via Spatiotemporal Quadric Element (Stequel) Matching

Mikhail Sizintsev
Richard P. Wildes

Department of Computer Science and Engineering
and the Centre for Vision Research
York University
Toronto, Ontario M3J 1P3
Canada

Revised

Abstract

Spatiotemporal stereo is concerned with recovery of the 3D structure of a dynamically changing scene from a sequence of stereo images. This paper aims at computing temporally coherent disparity maps without explicit recovery of motion. We make use of a spatiotemporal volume paradigm and consider extracted features we called **stequels** as basic matching primitives. The stequel matching principle is developed. Extensive algorithmic evaluation with ground truth data incorporated in both local and global correspondence paradigms shows the great benefit of considering stequels as a matching primitive that naturally incorporates local spatial and temporal structure and its advantages in comparison to alternative ways of enforcing temporal coherence in the stereo estimation procedure.

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Previous work	2
1.3	Contributions	3
2	Technical Approach	5
2.1	Spatiotemporal matching primitive	5
2.1.1	3D steerable filters	5
2.1.2	Building the match primitive	7
2.1.3	Spatiotemporal derivatives – Grammian	8
2.2	Spatiotemporal epipolar correspondence constraint	8
2.2.1	Left-Right Flow Relationship	10
2.2.2	General Left/Right Orientation Relationship	11
2.3	Stequel match cost	13
2.3.1	Subpixel matching	15
3	Empirical Evaluation	18
3.1	Algorithmic instantiations	18
3.2	<i>Lab</i> sequences	23
3.2.1	Stequel vs. Grammian	25
3.3	<i>Office</i> sequence	27
3.4	<i>Rover</i> sequence	27
4	Discussion	29
A	Flow recovery from stequel	31

List of Figures

2.1	Image Spacetime. xt slices for the left and the right views. . .	6
2.2	3D Steerable filter pair.	7
2.3	Stereo Geometry	9
3.1	<i>Lab1</i> Sequence Tests	19
3.2	<i>Lab2</i> Sequence Tests	20
3.3	Error statistics for the <i>Lab1</i> and <i>Lab2</i> Tests	21
3.4	<i>Office</i> Sequence Tests	22
3.5	<i>Rover</i> Sequence Tests	23
3.6	Error statistics for comparison of stequel vs. Grammian	26

Chapter 1

Introduction

1.1 Motivation

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

In response to the above observations, this paper describes a novel approach to recovering temporally coherent disparity estimates from a sequence of binocular images. The key idea is to base stereo correspondence on matching primitives that inherently encompass both the spatial and temporal dimensions of image spacetime. In particular, each temporal stream of imagery is locally represented in terms of its orientation structure, as captured by the spatiotemporal quadric (also variously referred to as the orientation tensor and covariance matrix, see, e.g., [1, 2]). By representing orientation structure uniformly across image space and time, both instantaneously defined (e.g., spatial texture) and dynamically defined (e.g., motion) information can be brought to bear on stereo correspondence in an integrated fashion. It will be shown that by basing matching on this representation, it is possible to

recover temporally coherent disparity estimates, without the need to make optical or 3D flow explicit. Although, extension of this work to allow for simultaneous estimates of disparity and flow is an interesting direction for future research. Further, this representation allows spatial and temporal image structure to resolve otherwise ambiguous matches in a fashion consistent with both sources of information. Significantly, applicability of this representation to stereo correspondence is quite general and will be demonstrated in both local and global matchers.

1.2 Previous work

Early work combining stereo and motion concentrated on punctate features (e.g., edges, corners). One of the earliest attempts made use of heuristics for assigning spatial and temporal matches based on model-based reasoning [3]. A rather different early approach exploited constraints on the temporal derivative of disparity [4]. Other work matched binocular features to recover 3D estimates for temporal tracking [5, 6]. More recent research that relies on loose coupling of stereo and motion has emphasized the recovery of 3D motion using optical flow in conjunction with multiple hypothesis binocular disparity maps [7]. The proposed research differs from such early work in being focused on a more integrated approach to spatiotemporal processing and in its emphasis on dense reconstruction (while [7] considered dense estimates, the stereo-motion coupling is loose).

More recent stereo research has seen increased interest in scene recovery from multi-camera (especially binocular) video as constrained by 3D models. Some work has concentrated on the recovery of surface mesh models between individual stereo pairs with tracking across time instances serving to yield temporally consistent models [8]. Other research considers multiple cameras, employs voxel carving for initial estimation and uses intensity-based matching over spatiotemporal volumes without consideration of image motion differences between different views [9]. Still other work casts stereo and motion estimation as a generic image matching problem solved variationally after backprojecting the input images onto a suitable surface [10]. Again, the proposed approach differs from these lines of research in its emphasis on a more integrated approach to stereo and motion and in eschewing explicit surface models, which can become problematic when dealing with multiple objects and complex scenes.

Other lines of recent research have emphasized more integrated approaches to stereo and motion processing. Some of this work has concentrated on static scenes with variable lighting [11]. Others have focused on defining appropriate temporal integration windows, e.g., as part of the correspondence process [12] or simply reinforce disparity estimates from the previous frame using optical flow [13]. Further, combined stereo and motion estimation has been formulated in terms of both PDEs [14, 15] as well as MRFs [16, 17, 18, 19, 20]. Still other work has used direct methods for integrated recovery of structure and egomotion [21, 22, 23]. The proposed research shares with these efforts an emphasis on tight integration of binocular imagery with time. It is novel in basing its matching on the representation of image spacetime in terms of local spatiotemporal orientation, which provides richer image descriptions than employed in previous methods, as they typically worked with raw image intensities.

A major tool that is employed in the proposed approach is the representation of spacetime imagery in terms of oriented spatiotemporal structure. Various research has documented optical flow recovery [24, 25], tracking [26] and grouping [27] on the basis of filters tuned for local spatiotemporal orientation. More specifically, previous research has considered the use of the spatiotemporal quadric to capture orientation in image spacetime, with application to motion estimation, restoration, enhancement [1, 2] and flow comparison [28]. However, it appears none has exploited spatiotemporal orientation, in general, or the spatiotemporal quadric, specifically, for stereo disparity estimation. Previous stereo work has defined binocular correspondence based on a bank of spatial filters [29]. The proposed approach also extracts its measures of orientation via application of a filter bank; however, it is significantly different in employing filters that span both the spatial and temporal domains, thereby basing matching on a fundamentally richer representation.

1.3 Contributions

In the light of previous research, the main contributions of this work are as follows. (i) The spatiotemporal quadric is proposed as a matching primitive for spacetime stereo. This primitive captures both local spatial and temporal structure and thereby enables matching to account for both sources of data without need to estimate optical flow or 3D motion. (ii) The geometric relationships between corresponding spatiotemporal quadrics across binocular

views are derived and used to motivate a match cost. The spatiotemporal match primitives and cost are incorporated in local and global matchers. (iii) Extensive empirical evaluation of these matchers is presented. Testing encompasses quantitative evaluation on laboratory acquired binocular video with ground truth and qualitative evaluation on more naturalistic imagery. The laboratory imagery and associated ground truth are available for download at <http://www.cse.yorku.ca/vision/research/ststereo.shtml>.

Chapter 2

Technical Approach

2.1 Spatiotemporal matching primitive

In dealing with temporal sequences of binocular images, it is possible to conceptualize of stereo correspondence in terms of image spacetime, which naturally encompasses both spatial and temporal characteristics of local pattern structure, see Fig. 2.1a. While image spacetime can be operated on directly, using pixel intensities, consideration of local spatiotemporal orientation provides access to a richer representation. Local orientation has visual significance as orientations parallel to the image plane capture the spatial pattern of observed surfaces (e.g., spatial texture); whereas, orientations that extend into the temporal dimension capture dynamic aspects (e.g., motion). By integrating the temporal dimension into the primitive, subsequent matching will be inherently constrained to observe temporal coherence. Further, through combination of both temporal and spatial structure in the descriptor, match ambiguities that might exist through consideration of only one data source have potential to be resolved.

2.1.1 3D steerable filters

To extract a representation of orientation from imagery, one can filter the data with oriented filters. In the current work, 3D Gaussian, second-derivative filters, G_2 , and their Hilbert transforms, H_2 [30], are applied to the data with responses pointwise rectified (squared) and summed. Filtering is executed across a set of 3D orientations given by unit column vectors, $\hat{\mathbf{w}}_i$. Hence a

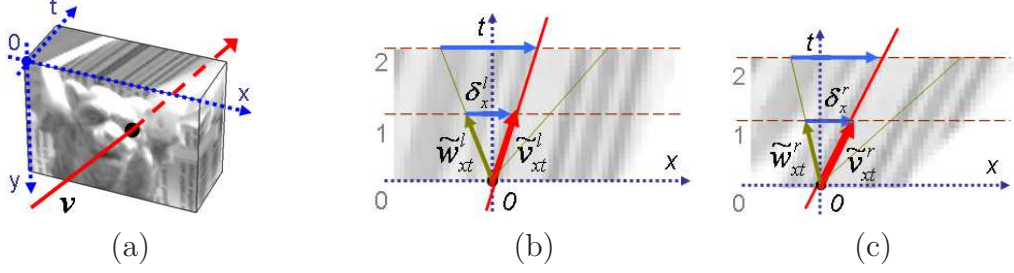


Figure 2.1: Image Spacetime. (a) Spacetime can be conceptualized as a spatiotemporal volume xyt . An instantaneous motion trajectory, v (shown in red), traces an orientation in this volume. (b) An exemplar xt slice of the spatiotemporal volume for the left view (c) The corresponding xt slice in the right view. \tilde{v}_{xt}^l and \tilde{v}_{xt}^r are the projections of the v^l and v^r onto the xt slice; w^l and w^r are arbitrary vectors (shown in green) in correspondence in xyt space and $\delta^r = \tilde{w}^r - \tilde{v}^r$, $\delta^l = \tilde{w}^l - \tilde{v}^l$ (shown in blue); $\delta^r = A\delta^l$ as explained in text.

measure of local energy, E , is computed according to

$$E(\mathbf{x}; \hat{\mathbf{w}}_i) = [G_2(\hat{\mathbf{w}}_i) * I(\mathbf{x})]^2 + [H_2(\hat{\mathbf{w}}_i) * I(\mathbf{x})]^2, \quad (2.1)$$

where $\mathbf{x} = (x, y, t)$ are spatiotemporal image coordinates, I is the image sequence and $*$ denotes convolution [30].

Figure 2.2 visualizes G_2 along a particular direction and its 90°-phase counterpart H_2 filter. The composed response as in (2.1) will be phase-invariant and specific to the chosen direction, which in practice results in better performance than typical spatiotemporal derivatives. Furthermore, even very oblique orientations which correspond to large motions can be sampled with G_2 - H_2 pairs, which is less reliable with simple local derivatives.

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

$$(\pm 1, \pm 1, \pm 1), (0, \pm 1/\phi, \pm \phi), (\pm 1/\phi, \pm \phi, 0), (\pm \phi, 0, \pm 1/\phi),$$

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

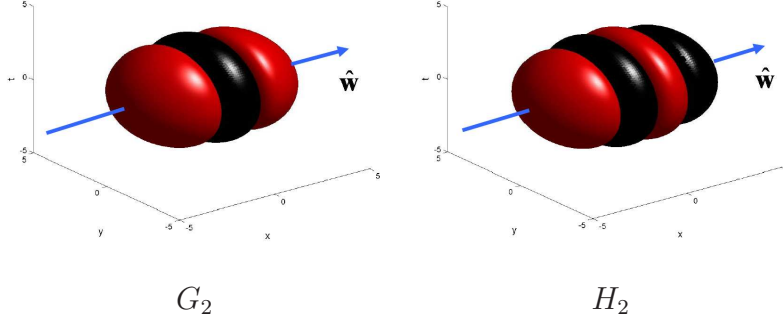


Figure 2.2: Surfaces plots of 3D steerable filter pair G_2 and H_2 oriented along the x -axis in the spacetime volume, i.e. $\hat{\mathbf{w}} = [1 \ 0 \ 0]^\top$. Red and black colours represent positive and negative contributions, respectively.

2.1.2 Building the match primitive

To proceed, the individual energy measures are recast in terms of the spatiotemporal quadric. This particular representation captures local orientation as well as the variance of spacetime about that orientation. This construct captures the local shape of spacetime (e.g., point- vs. line- vs. plane-like) in addition to direction for a local descriptor that is richer than if (dominant) orientation alone is considered [1]. Furthermore, the quadric casts structure in terms of spacetime coordinates, $\mathbf{x} = (x, y, t)$, where it is convenient to formulate binocular match constraints. In the context of binocular matching, this quadric will be referred to as the **stequel**, *spatio-temporal quadric element*, \mathbf{Q} . In particular,

$$\mathbf{Q} = \sum_i \hat{E}_i \hat{\mathbf{w}}_i \hat{\mathbf{w}}_i^\top, \quad (2.2)$$

where summation is across the set of filter orientations, $\hat{\mathbf{w}}_i$, and \hat{E}_i is the corresponding local energy response (2.1), but now normalized such that $\sum_i \hat{E}_i(\mathbf{x}) = 1$ (see Sec. 2.2, esp. footnote 1 for normalization rationale). In constructing \mathbf{Q} , the dyadic product, $\hat{\mathbf{w}}_i \hat{\mathbf{w}}_i^\top$, establishes the local frame implied by orientation $\hat{\mathbf{w}}_i$ weighted by its corresponding response, \hat{E}_i , [1].

For a binocular sequence, the stequel, \mathbf{Q} , is computed pointwise in space-time and separately for the left and right image sequences to provide matching primitives; thus, it is parametrized as $\mathbf{Q}^l(\mathbf{x})$ and $\mathbf{Q}^r(\mathbf{x})$, in reference to the left and right views, resp. Significantly, the implied calculations are modest.

The calculation of local energy is realized through steerable filters requiring nothing more than 3D separable convolution and pointwise nonlinearities and is thereby amenable to compact, efficient implementation [31]. Construction of \mathbf{Q} from the filter responses requires only matrix summation, as specified in (2.2). Nevertheless, depending on the observed efficacy of this particular filtering approach, alternatives may also be considered: As examples, Gabor and lognormal filters may be considered.

2.1.3 Spatiotemporal derivatives – Grammian

A simpler and computationally more efficient alternative to extract stequels is to construct quadric matching primitives based on spatiotemporal derivatives aggregated over some local region, also known as the Gram matrix, or Grammian, e.g., as used in [32] in a different context. In this case,

$$\mathbf{Q}_G = \frac{1}{k} \sum_{i \in \mathcal{W}} \nabla I_i (\nabla I_i)^\top = \frac{1}{k} \sum_{i \in \mathcal{W}} \begin{bmatrix} \frac{\partial I_i}{\partial x} \\ \frac{\partial I_i}{\partial y} \\ \frac{\partial I_i}{\partial t} \end{bmatrix} \begin{bmatrix} \frac{\partial I_i}{\partial x} & \frac{\partial I_i}{\partial y} & \frac{\partial I_i}{\partial t} \end{bmatrix}, \quad (2.3)$$

where \mathcal{W} is the local aggregation region and $k = \sum_{i \in \mathcal{W}} (\nabla I_i)^\top \nabla I_i$ is the normalization factor.

Although our subsequent development and experiments are based on steerable filters, stequels constructed from spatiotemporal derivatives (Grammians) may be very useful in practice, as they are easier and faster to compute. On the other hand, empirical comparisons in Sec. 3.2.1 show that stequels yield superior performance in application to spatiotemporal stereo matching.

2.2 Spatiotemporal epipolar correspondence constraint

In establishing correspondence between binocular sequences, it is incorrect simply to seek the most similar stequels, as local spatiotemporal orientation is expected to change between views due to the geometry of the situation. In this section, constraint is derived between corresponding stequels subject to rectified and otherwise calibrated binocular viewing. This constraint is derived in two steps. First, the relationship between local spatiotemporal

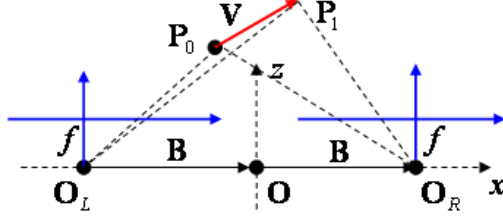


Figure 2.3: Stereo Geometry. A reference Euclidean coordinate system is centred at the midpoint of the stereo baseline, \mathbf{O} . Cameras are rectified with a half-baseline vector $\mathbf{b} = [b, 0, 0]^\top$ and focal lengths f . Left and right optical centres are at $O^l = -\mathbf{b}$ and $O^r = \mathbf{b}$, resp. Point \mathbf{P} undergoes an arbitrary displacement \mathbf{V} from instance 0 to 1.

orientations in left and right image spacetime is derived as a 3D scene point \mathbf{P} suffers an arbitrary (infinitesimal) 3D displacement, \mathbf{V} , relative to the imaging system. Here, displacement can come about through movement of the point, the imaging system or a combination thereof. Further, since the analysis is point-based, no scene rigidity is assumed.

While the relationship between left- and right-based flow has been investigated previously (e.g., [4]), the present derivation sets it in the light of left/right spatiotemporal orientation differences with application to disparity estimation; whereas, previous work assumed disparity estimation and was focused on subsequent 3D inferences. Further, the left/right flow relationships are generalized to capture the relationship between arbitrary orientations in left and right spacetimes. These results lead directly to the desired relationship between binocular stequels in correspondence.

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

2.2.1 Left-Right Flow Relationship

Consider how a 3D point, \mathbf{P} , is observed by the cameras as a function of time, t , while it is displaced along 3D direction, \mathbf{V} . The geometry of the situation is shown in Fig. 2.3. Cameras share a common intrinsic matrix

$$\mathbf{K} = \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

where other components of the matrix are accounted for by calibration and neglected. At time t , the projections of \mathbf{P} to the left and right views are given by

$$\begin{aligned} \mathbf{P}_t^l &= \mathbf{K}((\mathbf{P}_{t=0} - \mathbf{B}) + t\mathbf{V}) = \mathbf{P}_{t=0}^l + t\mathbf{K}\mathbf{V} \\ \mathbf{P}_t^r &= \mathbf{K}((\mathbf{P}_{t=0} + \mathbf{B}) + t\mathbf{V}) = \mathbf{P}_{t=0}^r + t\mathbf{K}\mathbf{V}. \end{aligned} \quad (2.4)$$

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

$$\mathbf{p}^l = \begin{bmatrix} p_x^l \\ p_y^l \end{bmatrix} = \begin{bmatrix} P_x^l/P_z^l \\ P_y^l/P_z^l \end{bmatrix} = \frac{1}{P_z^l} \begin{bmatrix} P_x^l \\ P_y^l \end{bmatrix} = Z^{-1}\mathbf{P}_{2 \times 1}^l, \quad (2.5)$$

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

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

$$\tilde{\mathbf{v}}^l = \begin{bmatrix} \mathbf{p}_{t=1}^l - \mathbf{p}_{t=0}^l \\ v_t^l \end{bmatrix} = \begin{bmatrix} \mathbf{p}_{t=1}^l - \mathbf{p}_{t=0}^l \\ 1 \end{bmatrix}, \quad (2.6)$$

where $v_t^l = 1$ by definition, as time has been taken from $t = 0$ to $t = 1$. Analogously for the right view

$$\tilde{\mathbf{v}}^r = \begin{bmatrix} \mathbf{p}_{t=1}^r - \mathbf{p}_{t=0}^r \\ 1 \end{bmatrix}. \quad (2.7)$$

To relate the left and right spatiotemporal orientations, it is useful to cast the left-camera flow vectors (2.6) and their right camera counterparts in

terms of temporally varying position (2.4) and (2.5). Left camera-based flow is given by (2.6) and substitution from (2.5) yields

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

Further substitution for \mathbf{P}^l according to (2.4) and letting all subscripts pertain to time (i.e, 0 and 1 denote $t = 0$ and $t = 1$, resp.) yields

$$\tilde{\mathbf{v}}_{2 \times 1}^l = \frac{Z_0 - Z_1}{Z_0 Z_1} \bar{\mathbf{K}} \mathbf{P}_0 + \frac{1}{Z_1} \bar{\mathbf{K}} \mathbf{V} - \frac{Z_0 - Z_1}{Z_0 Z_1} \bar{\mathbf{K}} \mathbf{B}, \quad (2.8)$$

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

$$\tilde{\mathbf{v}}_{2 \times 1}^r = \frac{Z_0 - Z_1}{Z_0 Z_1} \bar{\mathbf{K}} \mathbf{P}_0 + \frac{1}{Z_1} \bar{\mathbf{K}} \mathbf{V} + \frac{Z_0 - Z_1}{Z_0 Z_1} \bar{\mathbf{K}} \mathbf{B}. \quad (2.9)$$

Finally, the relationship between the left (2.8) and right (2.9) flows is revealed by taking their difference

$$\tilde{\mathbf{v}}^r - \tilde{\mathbf{v}}^l = \begin{bmatrix} 2(Z_0 - Z_1) \bar{\mathbf{K}} \mathbf{B} / (Z_0 Z_1) \\ 0 \end{bmatrix} = \begin{bmatrix} \Delta \\ 0 \\ 0 \end{bmatrix}, \quad (2.10)$$

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

2.2.2 General Left/Right Orientation Relationship

The relationship (2.10) was derived only for dominant motion orientation; whereas, stequels capture information from *all* directions $\tilde{\mathbf{w}}$ in (x, y, t) , which now are considered.

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

$$\delta^r = \tilde{\mathbf{w}}^r - \tilde{\mathbf{v}}^r = \begin{bmatrix} \delta_x^r & \delta_y^r & 0 \end{bmatrix}^\top, \quad (2.11)$$

$$\delta^l = \tilde{\mathbf{w}}^l - \tilde{\mathbf{v}}^l = \begin{bmatrix} \delta_x^l & \delta_y^l & 0 \end{bmatrix}^\top \quad (2.12)$$

that capture the purely spatial orientation of corresponding elements (see Fig. 2.1b,c). For the special case of fronto-parallel surfaces $\delta^r = \delta^l$, i.e. disregarding motion, oriented texture appears the same across binocular views.

For the more general case where surfaces are slanted with respect to the imaging system, the imaged orientation of corresponding elements changes across views, even in the absence of motion. For present matters, this change can be modeled by a linear transformation $\delta^r = \mathbf{A}\delta^l$. Considering that the third element of the δ vectors is always zero by construction, and $\delta_y^r = \delta_y^l$ due to conventional stereo epipolar constraints for rectified setups, this relationship takes the form

$$\delta^r = \mathbf{A}\delta^l, \text{ where } \mathbf{A} = \begin{bmatrix} a_1 & a_2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \quad (2.13)$$

Substituting (2.11), (2.12) into (2.13) and rearranging yields,

$$\tilde{\mathbf{w}}^r = \mathbf{A}\tilde{\mathbf{w}}^l - \mathbf{A}\tilde{\mathbf{v}}^l + \tilde{\mathbf{v}}^r. \quad (2.14)$$

Further substitution of (2.10) results in

$$\begin{aligned} \tilde{\mathbf{w}}^r &= \mathbf{A}\tilde{\mathbf{w}}^l + \left(-\mathbf{A}\tilde{\mathbf{v}}^l + \tilde{\mathbf{v}}^l + \begin{bmatrix} \Delta & 0 & 0 \end{bmatrix}^\top \right) \\ &= \begin{bmatrix} a_1 & a_2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tilde{\mathbf{w}}^l + \begin{bmatrix} 1-a_1 & -a_2 & \Delta \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tilde{\mathbf{v}}^l \\ &= \begin{bmatrix} a_1 & a_2 & ((1-a_1)\tilde{v}_x^l - a_2\tilde{v}_y^l + \Delta) \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tilde{\mathbf{w}}^l \end{aligned} \quad (2.15)$$

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

$$\tilde{\mathbf{w}}^r = \mathbf{H}\tilde{\mathbf{w}}^l, \text{ where } \mathbf{H} = \begin{bmatrix} 1+h_1 & h_2 & h_3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (2.16)$$

It is interesting to outline a special case associated with (2.16). The situation of $h_1 = h_2 = 0$ means that $\delta^r = \delta^l$ in (2.13), which essentially implies the fronto-parallel assumption, that is still widely used in contemporary stereo matching. This case is quite important from a practical point of view, because it yields reasonable results and is faster as well as more numerically stable in estimation owing to its simpler form.

With (2.16) in place, it is possible to relate corresponding stequels. By design, (2.2), stequel **Q** reveals the amount of intensity variation along all

directions in spacetime, and the response ϕ to unit direction $\hat{\mathbf{w}} = \mathbf{w}/\sqrt{\mathbf{w}^T \mathbf{w}}$ is

$$\phi = \hat{\mathbf{w}}^T \mathbf{Q} \hat{\mathbf{w}}, \quad (2.17)$$

see, e.g., [1]. Assuming that spatiotemporal correspondences vary in orientation pattern, but not in the intensity per se¹, the responses, ϕ^l, ϕ^r , of corresponding stequels, $\mathbf{Q}^l, \mathbf{Q}^r$, must be the same for related directions, $\hat{\mathbf{w}}^l, \hat{\mathbf{w}}^r$, i.e.

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

Expanding the normalizations of $\hat{\mathbf{w}}^l$ and $\hat{\mathbf{w}}^r$ and substituting from (2.16) produces

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

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

$$\frac{\hat{\mathbf{w}}^{l\top} \mathbf{Q}^l \hat{\mathbf{w}}^l}{\hat{\mathbf{w}}^{l\top} \hat{\mathbf{w}}^l} = \frac{\hat{\mathbf{w}}^{l\top} \mathbf{H}^\top \mathbf{Q}^r \mathbf{H} \hat{\mathbf{w}}^l}{\hat{\mathbf{w}}^{l\top} \mathbf{H}^\top \mathbf{H} \hat{\mathbf{w}}^l}. \quad (2.18)$$

Since (2.18) holds for arbitrary orientations $\hat{\mathbf{w}}^l$ when \mathbf{Q}^l and \mathbf{Q}^r are stequels in correspondence, it provides the sought for general constraint on binocular stequels. It will be referred to as the *stequel correspondence constraint* and used to derive an approach to stereo matching.

2.3 Stequel match cost

To determine whether two stequels $\mathbf{Q}^l(x, y, t)$ and $\mathbf{Q}^r(x + d, y, t)$ are in correspondence with disparity d , a match cost must be defined. In this section, this cost is derived based on the stequel correspondence constraint, (2.18), and is taken as the error residual that results from solving for $\mathbf{h} = [h_1 \ h_2 \ h_3]^\top$ given two candidate stequels.

For a given direction vector $\hat{\mathbf{w}}_m^l$ at some particular orientation m and matching stequels, \mathbf{Q}^l and \mathbf{Q}^r , the stequel correspondence constraint, (2.18), yields a quadratic equation in the unknowns of \mathbf{h} of the form

$$\begin{aligned} f_m(\mathbf{h}) &= (\hat{\mathbf{w}}_m^{l\top} \mathbf{Q}^l \hat{\mathbf{w}}_m^l) (\hat{\mathbf{w}}_m^{l\top} \mathbf{H}^\top \mathbf{H} \hat{\mathbf{w}}_m^l) \\ &\quad - (\hat{\mathbf{w}}_m^{l\top} \hat{\mathbf{w}}_m^l) (\hat{\mathbf{w}}_m^{l\top} \mathbf{H}^\top \mathbf{Q}^r \mathbf{H} \hat{\mathbf{w}}_m^l) = 0. \end{aligned} \quad (2.19)$$

¹This is a weak form of brightness constancy as any additive and multiplicative intensity offsets between correspondences are compensated for by the bandpass and normalized filters used in stequel construction (2.2).

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

$$E_4 = \sum_{m=1}^M f_m(\mathbf{h})^2, \quad (2.20)$$

which is quartic in the entries of \mathbf{h} . While such a solution could be sought through analytic or numerical means, it has potential to be expensive to compute and noise sensitive owing to its order. Therefore, it is useful to linearize each error Eqn. (2.19) through expansion as a Taylor series in \mathbf{h} and retention of terms only through first-order to get

$$g_m(\mathbf{h}) = f_m(\mathbf{0}) + \nabla f_m^\top(\mathbf{0})\mathbf{h}, \quad (2.21)$$

with $\mathbf{0}$ being the $M \times 1$ zero vector. Using (2.21), the final function to be minimized with respect to \mathbf{h} becomes

$$E_2 = \sum_{m=1}^M (f_m(\mathbf{0}) + \nabla f_m^\top(\mathbf{0})\mathbf{h})^2, \quad (2.22)$$

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

$$\mathbf{G} = [\nabla f_1^\top(\mathbf{0}), \nabla f_2^\top(\mathbf{0}), \dots, \nabla f_M^\top(\mathbf{0})]^\top$$

and

$$\mathbf{c} = -[f_1(\mathbf{0}), f_2(\mathbf{0}), \dots, f_M(\mathbf{0})]^\top$$

yields

$$\begin{aligned} \mathbf{h} &= (\mathbf{G}^\top \mathbf{G})^{-1} \mathbf{G}^\top \mathbf{c}; \\ E_2 &= \|\mathbf{G}\mathbf{h} - \mathbf{c}\|_2^2 = \mathbf{c}^\top \mathbf{c} - (\mathbf{G}^\top \mathbf{c})^\top (\mathbf{G}^\top \mathbf{G})^{-1} \mathbf{G}^\top \mathbf{c}. \end{aligned} \quad (2.23)$$

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

Significantly, preliminary experiments showed that match cost based on the linearized error, (2.22), yielded slightly superior results to considering the original nonlinear error, (2.20), which lends further support to pursuing the advocated (linearized) approach. This could be partially explained by the

fact that interest is in a *discriminative* error measure that reliably penalizes bad matches, and not in the precise error value per se. Furthermore, linearized solution, along with being straightforward to compute, is also more robust in practice, as solutions to the higher-order equations are inherently more sensitive to noise. Here, noise arises via standard image corruptions in the input left and right image datastreams. Since the stequels are derived from the left and right image streams, they also are corrupted; thereby, the match cost computation that operates over the stequels will be subject to noise. Finally, since matching must be done for every point, it must be sufficiently simple to be practical - solution (2.23) requires the inverse of a 3x3 matrix, but it can be coded in closed form; indeed, the whole matching procedure is comparable to normalized cross correlation in terms of computational complexity and runtime.

2.3.1 Subpixel matching

We have previously described the matching constraint between left and right stequels separated by a particular integer disparity value. To get subpixel disparity we adapt the standard Lucas-Kanade technique [33, 34] to stequel-based stereo. As before, the objective is to minimize the mean-squared error ρ over subpixel displacement, $\vec{\tau}$, in the form

$$\rho(\vec{\tau}) = \sum_{m,i} [\hat{\mathbf{w}}_m^\top \mathbf{Q}_i^l \hat{\mathbf{w}}_m \hat{\mathbf{w}}_m^\top \mathbf{H}^\top \mathbf{H} \hat{\mathbf{w}}_m - \hat{\mathbf{w}}_m^\top \mathbf{H}^\top \mathbf{Q}_i^r [\mathcal{W}(\vec{\tau})] \mathbf{H} \hat{\mathbf{w}}_m]^2 \quad (2.24)$$

where m is the index over the spanning set of 3D directions, i is the index over spatial aggregation (if such exists), which we suppress in the following derivations for convenience, and $\mathcal{W}(\tau)$ is the warping function.

We approximate $\mathbf{Q}^r [\mathcal{W}(\vec{\tau})]$ via the first order Taylor expansion as

$$\mathbf{Q} [\mathcal{W}(\tau)] = \mathbf{Q}(0) + \nabla \mathbf{Q} \vec{\tau}. \quad (2.25)$$

Since the stereo frames were captured at the same time and rectified imagery is assumed, we can safely use the restriction $\vec{\tau} = [\tau \ 0 \ 0]^\top$, which results in

$$\nabla \mathbf{Q} \vec{\tau} = \tau \frac{\partial \mathbf{Q}}{\partial \tau} = \tau \begin{bmatrix} \frac{\partial q_{11}}{\partial x} & \frac{\partial q_{12}}{\partial x} & \frac{\partial q_{13}}{\partial x} \\ \frac{\partial q_{12}}{\partial x} & \frac{\partial q_{22}}{\partial x} & \frac{\partial q_{23}}{\partial x} \\ \frac{\partial q_{13}}{\partial x} & \frac{\partial q_{23}}{\partial x} & \frac{\partial q_{33}}{\partial x} \end{bmatrix} = \tau \mathbf{Q}_x \quad (2.26)$$

Thus, our objective is to find τ_0 such that

$$\begin{aligned}\tau_0 &= \arg \min_{\tau} \sum [\hat{\mathbf{w}}^\top \mathbf{Q}^l \hat{\mathbf{w}} \hat{\mathbf{w}}^\top \mathbf{H}^\top \mathbf{H} \hat{\mathbf{w}} - \hat{\mathbf{w}}^\top \mathbf{H}^\top \mathbf{Q}^r [\mathcal{W}(\tau)] \mathbf{H} \hat{\mathbf{w}}]^2 \\ &= \arg \min_{\tau} \sum [\hat{\mathbf{w}}^\top \mathbf{Q}^l \hat{\mathbf{w}} \hat{\mathbf{w}}^\top \mathbf{H}^\top \mathbf{H} \hat{\mathbf{w}} - \hat{\mathbf{w}}^\top \mathbf{H}^\top (\mathbf{Q}^r + \tau \mathbf{Q}_x^r) \mathbf{H} \hat{\mathbf{w}}]^2\end{aligned}\quad (2.27)$$

Since (2.27) is quadratic in terms of τ , the minimum is readily calculated analytically:

$$\tau_0 = \frac{\sum [(\hat{\mathbf{w}}^\top \mathbf{Q}^l \hat{\mathbf{w}} \hat{\mathbf{w}}^\top \mathbf{H}^\top \mathbf{H} \hat{\mathbf{w}} - \hat{\mathbf{w}}^\top \mathbf{H}^\top \mathbf{Q}^r \mathbf{H} \hat{\mathbf{w}}) \hat{\mathbf{w}}^\top \mathbf{H}^\top \mathbf{Q}_x^r \mathbf{H} \hat{\mathbf{w}}]}{\sum (\hat{\mathbf{w}}^\top \mathbf{H}^\top \mathbf{Q}_x^r \mathbf{H} \hat{\mathbf{w}})^2} \quad (2.28)$$

$$\begin{aligned}\rho(\tau_0) &= \sum (\hat{\mathbf{w}}^\top \mathbf{Q}^l \hat{\mathbf{w}} \hat{\mathbf{w}}^\top \mathbf{H}^\top \mathbf{H} \hat{\mathbf{w}} - \hat{\mathbf{w}}^\top \mathbf{H}^\top \mathbf{Q}^r \mathbf{H} \hat{\mathbf{w}})^2 - \\ &\quad \frac{(\sum [(\hat{\mathbf{w}}^\top \mathbf{Q}^l \hat{\mathbf{w}} \hat{\mathbf{w}}^\top \mathbf{H}^\top \mathbf{H} \hat{\mathbf{w}} - \hat{\mathbf{w}}^\top \mathbf{H}^\top \mathbf{Q}^r \mathbf{H} \hat{\mathbf{w}}) \hat{\mathbf{w}}^\top \mathbf{H}^\top \mathbf{Q}_x^r \mathbf{H} \hat{\mathbf{w}}])^2}{\sum (\hat{\mathbf{w}}^\top \mathbf{H}^\top \mathbf{Q}_x^r \mathbf{H} \hat{\mathbf{w}})^2}\end{aligned}\quad (2.29)$$

The expression (2.29) can be used as the (subpixel) matching error, but

first we must find the best \mathbf{H} of the form $\begin{bmatrix} 1 + h_1 & h_2 & h_3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ (as in (2.16))

that yields a minimum. So, to proceed, we find the $\mathbf{h} = [h_1 \ h_2 \ h_3]^\top$ that minimizes (2.29), which happens when the first derivatives with respect to h_j , $j \in \{1, 2, 3\}$ equal to zero, i.e. the gradient with respect to \mathbf{h} must be the zero vector. In particular,

$$\begin{aligned}\xi_j &= \frac{\partial \rho(\tau_0)}{\partial h_j} = \frac{\partial \sum (\hat{\mathbf{w}}^\top \mathbf{Q}^l \hat{\mathbf{w}} \hat{\mathbf{w}}^\top \mathbf{H}^\top \mathbf{H} \hat{\mathbf{w}} - \hat{\mathbf{w}}^\top \mathbf{H}^\top \mathbf{Q}^r \mathbf{H} \hat{\mathbf{w}})^2}{\partial h_j} \\ &\quad - \frac{\frac{\partial \sum [(\hat{\mathbf{w}}^\top \mathbf{Q}^l \hat{\mathbf{w}} \hat{\mathbf{w}}^\top \mathbf{H}^\top \mathbf{H} \hat{\mathbf{w}} - \hat{\mathbf{w}}^\top \mathbf{H}^\top \mathbf{Q}^r \mathbf{H} \hat{\mathbf{w}}) \hat{\mathbf{w}}^\top \mathbf{H}^\top \mathbf{Q}_x^r \mathbf{H} \hat{\mathbf{w}}]}{\partial h_j} \sum (\hat{\mathbf{w}}^\top \mathbf{H}^\top \mathbf{Q}_x^r \mathbf{H} \hat{\mathbf{w}})^2}{\left[\sum (\hat{\mathbf{w}}^\top \mathbf{H}^\top \mathbf{Q}_x^r \mathbf{H} \hat{\mathbf{w}})^2 \right]^2} \\ &\quad + \frac{\frac{\partial \sum (\hat{\mathbf{w}}^\top \mathbf{H}^\top \mathbf{Q}_x^r \mathbf{H} \hat{\mathbf{w}})^2}{\partial h_j} \sum [(\hat{\mathbf{w}}^\top \mathbf{Q}^l \hat{\mathbf{w}} \hat{\mathbf{w}}^\top \mathbf{H}^\top \mathbf{H} \hat{\mathbf{w}} - \hat{\mathbf{w}}^\top \mathbf{H}^\top \mathbf{Q}^r \mathbf{H} \hat{\mathbf{w}}) \hat{\mathbf{w}}^\top \mathbf{H}^\top \mathbf{Q}_x^r \mathbf{H} \hat{\mathbf{w}}]}{\left[\sum (\hat{\mathbf{w}}^\top \mathbf{H}^\top \mathbf{Q}_x^r \mathbf{H} \hat{\mathbf{w}})^2 \right]^2} \\ &= 0\end{aligned}\quad (2.30)$$

Since $\xi_j = 0$ is a polynomial of high order, we solve the first-order approximation to this system of equations, i.e.

$$\xi(\mathbf{h}) = \xi(\mathbf{0}) + \nabla \xi(\mathbf{0})^\top \mathbf{h} \quad (2.31)$$

Once \mathbf{h}_0 that satisfies $\xi(\mathbf{h}) = 0$ has been found by solving the corresponding linear equations (similar to (2.23)), this value is used directly in (2.28) and (2.29) to get the subpixel disparity and corresponding error measure, respectively. This procedure is subject to repetition several times for better convergence to the final and more precise result.

Chapter 3

Empirical Evaluation

3.1 Algorithmic instantiations

A software implementation has been developed that inputs a binocular video, computes stequels $Q^l(x, y, t)$ and $Q^r(x, y, t)$ for both sequences according to formula (2.2) and calculates the local match cost, (2.23), for any given disparity d , i.e., for stequels related as $Q^l(x, y, t)$ and $Q^r(x + d, y, t)$. To show the applicability of this approach to disparity estimation, the local match cost, (2.23), has been embedded in a coarse-to-fine local block-matching algorithm with shiftable windows [35] working over a Gaussian pyramid and also in a global graph-cuts with occlusions matcher [36] operating at the finest scale only; these matchers will be denoted **ST-local** and **ST-global**. Pixel-based disparity estimates are brought to subpixel precision via a Lucas-Kanade type refinement for stequels, as explained in Sec. 2.3.1.

To compare with non-stequel matching, versions of the local and global matchers that work simply on single left/right frame pixel comparisons are considered; these matchers will be denoted **noST-local** and **noST-global**, resp. Here, the normalized cross-correlation was used as the data cost term for local and global matching. Finally, to compare to an alternative method for enforcing temporal coherence, optical flow is estimated and used to define a spatiotemporal direction for match cost aggregation that operates over an equivalent number of frames as does the oriented filtering used in stequel construction (2.1). Here, optical flow is recovered from the stequel representation itself (see Appendix A and [1] for discussion) to make the comparison fair. The optical flow-based temporal aggregation is used only in conjunction with

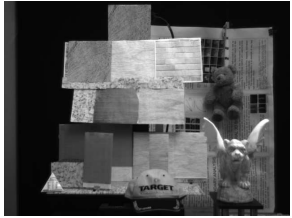
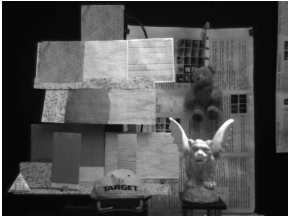






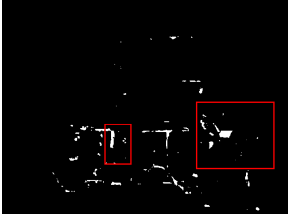
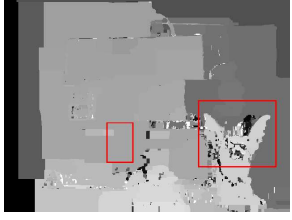
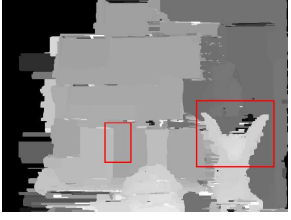
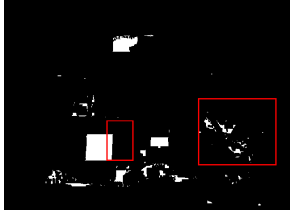

Lab 1 Left frame 12	Lab 1 Right frame 12	GT disparity
		
noST-local disparity	flowAg-local disparity	ST-local disparity
		
noST-local error	flowAg-local error	ST-local error
		
noST-global disparity		ST-global disparity
		
noST-global error		ST-global error
		

Figure 3.1: *Lab1* Tests. Example left and right frame 12 (out of 28 frames) with ground truth disparity (top row). Labeled boxes (beneath) show recovered disparity maps for compared algorithms and disparity-ground truth absolute differences. A few regions of particular interest in comparing results are highlighted with red rectangles, best seen in color. See accompanying videos at <http://www.cse.yorku.ca/vision/research/ststereo.shtml>



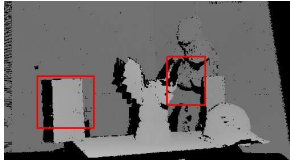



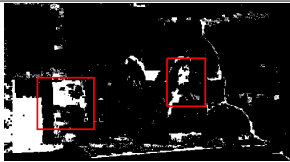
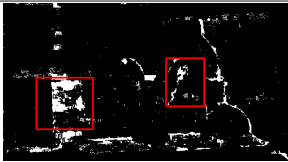
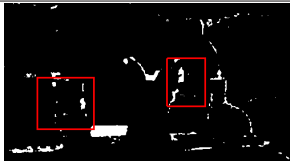
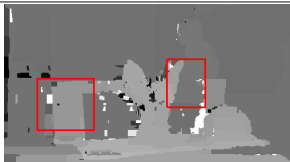
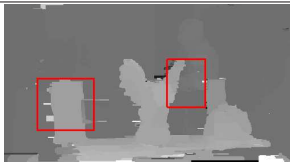
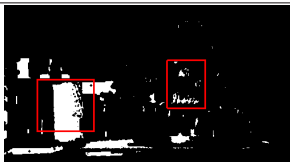
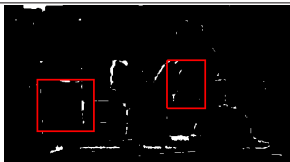
Lab 2 Left frame 10	Lab 2 Right frame 10	GT disparity
		
noST-local disparity	flowAg-local disparity	ST-local disparity
		
noST-local error	flowAg-local error	ST-local error
		
noST-global disparity		ST-global disparity
		
noST-global error		ST-global error
		

Figure 3.2: *Lab2* Tests. Example left and right frame 10 (out of 40 frames) with ground truth disparity (top row). Labeled boxes (beneath) show recovered disparity maps for compared algorithms and disparity-ground truth absolute differences. A few regions of particular interest in comparing results are highlighted with red rectangles, best seen in color. See accompanying videos at <http://www.cse.yorku.ca/vision/research/ststereo.shtml>

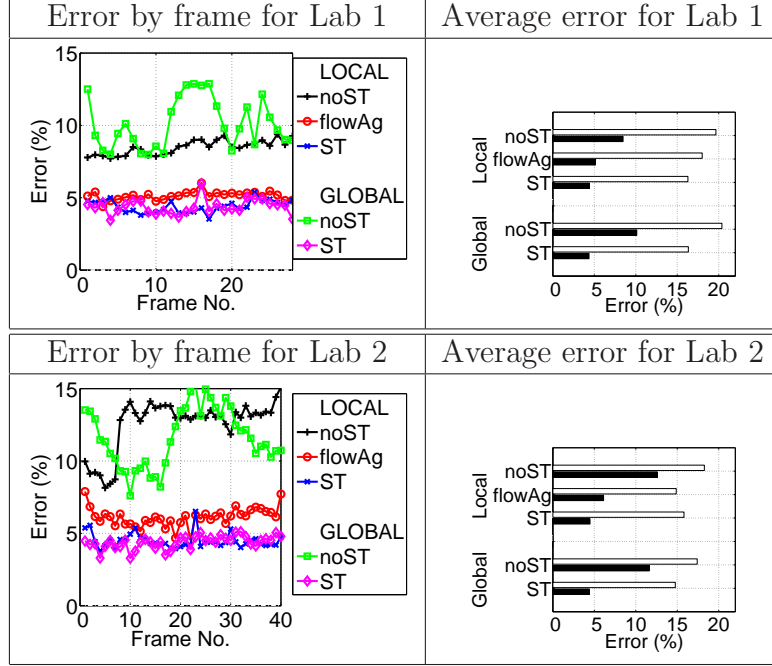


Figure 3.3: Error statistics for the *Lab1* and *Lab2* Tests. An error is taken as greater than 1 pixel discrepancy between estimated and groundtruth disparity. Bar plots show average error across entire sequences: White bars are for points within 5 pixels of a surface discontinuity; black bars show overall error. Error by frame plots show percentage of points in error overall for each frame separately

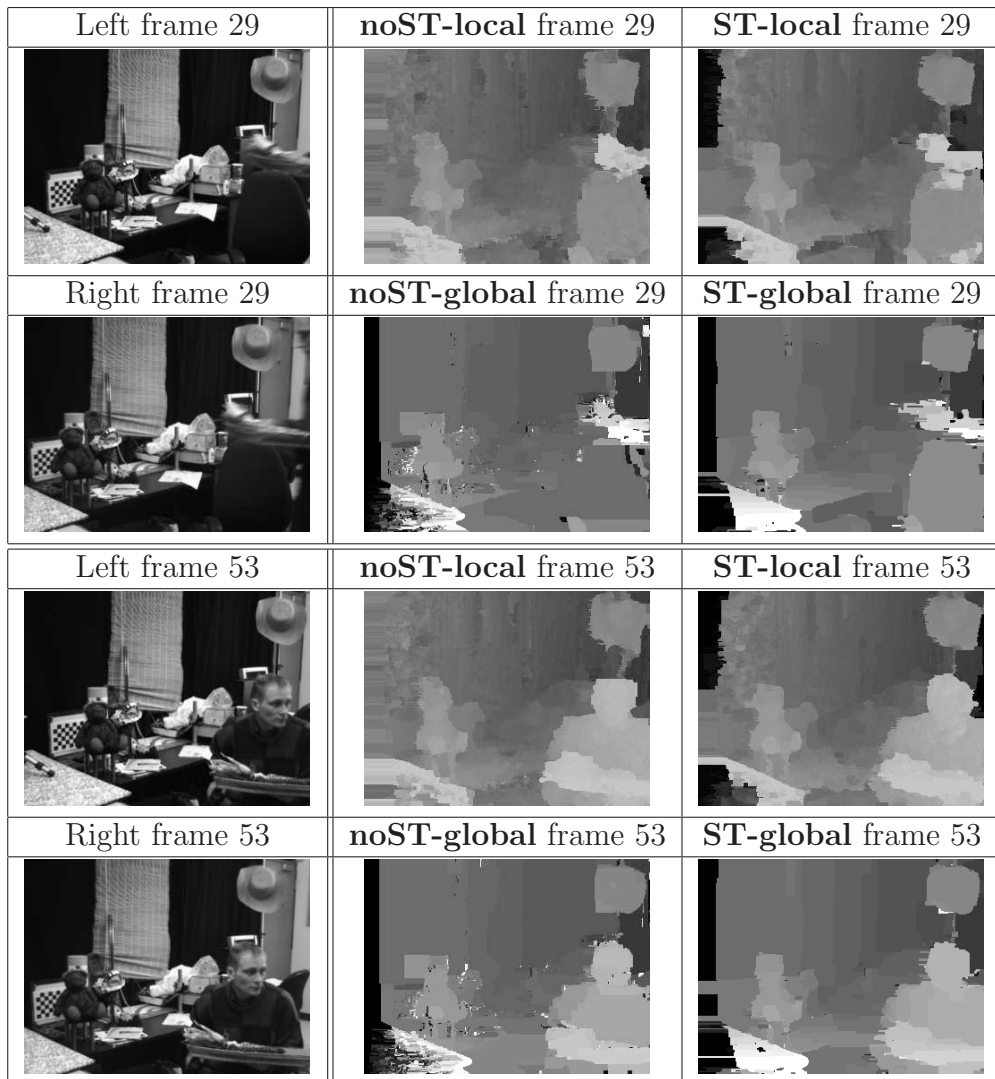


Figure 3.4: *Office* Tests. Right column shows left and right images for frames 29 and 53. Remaining boxes are labeled with recovered disparity by algorithm and frame. See accompanying videos at <http://www.cse.yorku.ca/vision/research/ststereo.shtml>

the local matcher, as incorporation into the global matcher by constructing a spatiotemporal MRF graph [37] is beyond the scope of this paper. The local flow-based aggregation matcher will be denoted **flowAg-local**.

In general, the comparison of local methods is important, as their results







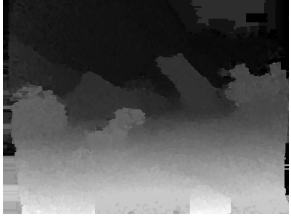

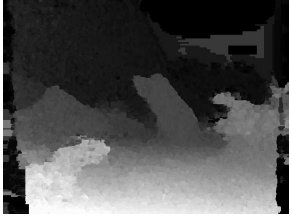
Left frame 11	Left frame 28	Left frame 53
		
flowAg-local frame 11	flowAg-local frame 28	flowAg-local frame 53
		
ST-local frame 11	ST-local frame 28	ST-local frame 53
		

Figure 3.5: *Rover Tests*. Top row shows left view at frames 88, 105 and 130. Recovered disparity maps at corresponding times are shown below for two algorithms. See accompanying videos at <http://www.cse.yorku.ca/vision/research/ststereo.shtml>

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

3.2 *Lab* sequences

Two laboratory data sets are considered. The first is a sequence (*Lab1*) captured with BumbleBee stereo camera [38] with (framewise) ground truth

disparity and discontinuity maps recovered according to a well-known structured light approach [39], see Fig. 3.1. This scene includes planes slanted in depth with texture oriented along epipolar lines (upper-central part of the scene), various bar-plane arrangement with identical repetitive textures (lower-central part of the scene) and complicated objects with non-trivial 3D boundaries and non-Lambertian materials (e.g., the teddy bear and gargoyle). For this sequence the stereo camera makes a complicated motion that translates along horizontal and depth axes, while part of the scene moves up and down; both camera and scene are on motorized stages.

Visual inspection of the image results (Fig. 3.1) shows that **noST-local** performs relatively poorly. Planar regions with epipolar aligned texture are generally difficult. Simple temporal aggregation provided by **flowAg-local** is seen to improve on these difficulties; however, performance degrades near 3D boundaries due to unreliable recovery of flow estimates in such areas. **ST-local** does the best of the three local matchers as its *ability to include temporal information allows it to resolve match ambiguities without explicit flow recovery*. As particular improvements of **ST-local** over **noST-local** and **flowAg-local**, consider the lower right and left regions marked with red rectangles in Fig. 3.1, which highlight the complex outline of the gargoyle wings and the vertical bar in front of plane both having identical textures (camouflage). **ST-local** is quite accurate in these challenging regions, while the other local methods perform relatively poorly. Objects located at different depths in space give rise to different image motions, even if they undergo the same world motion – and this difference is captured with stequels not allowing for improper matches.

For the global matchers, it is seen even with **noST-global** that it is possible to recover more precisely the complicated 3D boundaries and to achieve good disparity estimates in low texture regions via propagation from better defined boundary matches. However, **noST-global** performs poorly in the regions with epipolar aligned texture and camouflage, as initially incorrect estimates are not subsequently corrected. While increasing the smoothness improves on epipolar-aligned textures, it comes at the expense of camouflage resolution and vice versa. In comparison, **ST-global** is able to recover disparity reliably in these regions, as *once again the stequel representation supports proper resolution of situations that are ambiguous from the purely spatial information*. Another apparent advantage of the **ST-global** is more temporally consistent results – occasional mismatches in **noST-global** can be significantly amplified by propagating into nearby regions.

A second lab sequence, *Lab2*, is constructed in the same controlled environment as *Lab1*, but acquired with significant depth motion and out-of-plane rotation. This particular motion configuration is the most difficult for spatiotemporal stereo, as it results in significantly different left and right spatiotemporal volumes due to slanted surfaces and depth motion. Furthermore, large image motions are present in the individual left and right sequences. Figure 3.2 presents sample frame results for all five algorithmic instantiations considered above. Here, the conclusions reached from the analysis of *Lab1* are reinforced. With respect to the local methods, **ST-local** provides the most benefit both in weakly textured regions and near 3D boundaries. The performance of **flowAg-local** is hampered by large image motions, which are problematic to recover explicitly in this case; whereas, *direct stequel-based matching is still able to capitalize on temporal information without resolving flow and thereby operates well in the presence of nontrivial motions*. With respect to the global methods, the stequel-based matching **ST-global** significantly outperforms its pixel-based counterpart **noST-global**, especially for weakly-textured highly slanted foreground surfaces.

Error plots for both *Lab1* and *Lab2* quantify the improvements of stequel-based matching in comparison to rivals **noST** and **flowAg** (Fig. 3.3). Average errors across the sequences show the benefit of stequels near discontinuities and overall for both local and global matchers. Plots of error/frame reinforce the average improvements, but also document improved temporal coherence, as the stequel-based plots vary relatively little across frames, especially in comparison to purely spatial matching provided by **noST**. Incorporation of the temporal dimension also benefits **flowAg**, as its frame-by-frame statistics are relatively stable (albeit overall inferior to stequels); however, the more naturalistic imagery of the following examples further emphasizes the superior temporal coherence offered by stequels, even in comparison to **flowAg**.

3.2.1 Stequel vs. Gramian

In this section, two versions of **ST-local** have been compared on the *Lab1* and *Lab2* test sets. One version made use of stequel matching primitives, as described in Sec. 2.1, and the other made use of Gramian matching primitives (Sec. 2.1.3, as an alternative (denoted here as **Gram**). For the latter, spatiotemporal gradients were computed using optimized gradient filters [40].

Figure 3.6 shows quantitative results of **3Dfilt** and **Gram** on *Lab1* and

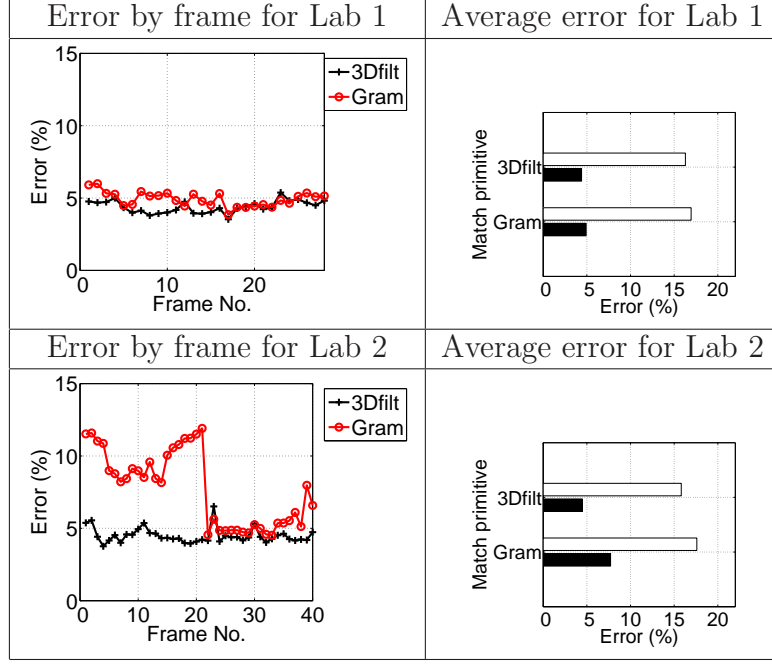


Figure 3.6: Error statistics for the *Lab1* and *Lab2*. Tests for the instantiations of **ST-local** operating on stequels constructed from spatiotemporal energies (**3Dfilt**) and Gram matrix constructed from the first-order spatiotemporal derivatives (**Gram**). An error is taken as greater than 1 pixel discrepancy between estimated and groundtruth disparity. Bar plots show average error across entire sequences: White bars are for points within 5 pixels of a surface discontinuity; black bars show overall error. Error by frame plots show percentage of points in error overall for each frame separately.

Lab2. The results document the superior performance realized by the matching primitives based on Gaussian derivative energy filters, **3Dfilt**, in comparison to matching primitives based on simpler spatiotemporal derivatives, **Gram**. Specifically, stequels exhibit more stable behaviour and yield consistently lower errors. The differences are particularly pronounced in the results for *Lab2*, where the image motions are relatively large and **Gram** yields occasional gross errors. These differences in performance may be accounted for by the finer orientational tuning that is offered by the energy filters, as well as their phase invariant responses.

3.3 *Office* sequence

The third data set, *Office*, depicts a more naturalistic (albeit without ground truth) cluttered indoor office scene where the camera pans while a person enters and subsequently moves about in a nonrigid fashion, see Fig. 3.4. Here, the superior ability of stequel matching to produce temporally coherent disparity maps is illustrated, as both **ST-local** and **ST-global** best their non-stequel-based counterparts. While temporal coherence is appreciated most by viewing the corresponding videos, observations can be made with respect to Fig. 3.4. For example, notice the more consistent disparity estimates recovered for the low texture walls and the chair via stequel matching, the lack of sudden, high variation, seen both with **noST-local** and **noST-global**, and the more accurate outlines of the teddy bear, the head and the hat suspended above.

3.4 *Rover* sequence

The fourth data set, *Rover*, is an outdoor sequence acquired from a robot rover traversing rugged terrain, including a receding foreground plane, a central diagonal rock outcropping, left side cliff, various boulders and bushes.

For this case, prior to processing with the stereo algorithms, the sequence was stabilized in software to compensate for the extremely jerky camera motion: Stabilization operated by warping neighboring frames to reference frames throughout the video according to affine transformations recovered via a parametric motion estimator [41]. For presentation, however, results are shown with respect to the original (unstabilized) video.

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

Chapter 4

Discussion

This paper described a novel approach to recovering temporally coherent disparity estimates using stequels as a spatiotemporal matching primitive. Temporal coherence arises naturally, as the primitives and the match cost inherently involve the temporal dimension. Further, matches that are ambiguous when considering only spatial pattern are resolved through the inclusion of temporal information. The stequel matching machinery is simple and involves linear computations only, (2.23). Thorough experimental evaluation on various datasets shows the benefit of stequel matching as incorporated both in local and global algorithms. Stereo sequences with ground truth have been introduced and are available online for comparison with other algorithms.

A particularly notable benefit of stequel matching is the ability to incorporate temporal information *without* image motion recovery. Optical flow estimation is challenging near 3D boundaries, weakly-textured regions and susceptible to an aperture problem – importantly, this paper demonstrated that stequels are powerful in exactly these situations and provide truly temporally coherent estimates with fewer isolated gross errors. Apparently, stequels allow stereo matching to capitalize on available spatiotemporal structure, even when optical flow recovery is difficult. By necessarily committing to local flow vectors, especially when data is insufficient for such interpretation, optical flow yields unreliable temporal aggregation; in contrast, stequels more completely characterize whatever spatiotemporal structure is present and make it available for appropriate matching. Further, note that it is non-trivial to model continuity in time with, e.g., an MRF prior model as, strictly speaking, temporal graph links have to be defined by flow (as in [20]). Ste-

quels, on the other hand, are directly applicable to standard 2D MRF graphs and their successful performance has been documented in this paper.

In conclusion, a computationally tractable and simple solution to spatiotemporal stereo has been presented, which proved to be very reliable, versatile and robust in practice. Significantly, this is the first attempt at stequel-based matching and various extensions can be considered, e.g., exploiting the spatiotemporal profile for explicit non-Lambertian and multi-layer matching. Also, extensions to 3D motion recovery can be considered using stequels in correspondence, which is anticipated to be beneficial, as stequels allow for *simultaneous* analysis of the temporal pattern in both views in addition to 3D structure estimation.

Acknowledgements

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

Appendix A

Flow recovery from stequel

The recovery of optical flow from stequels is outlined briefly in this appendix. Details can be found in [1].

Let \mathbf{Q} be a stequel at $\mathbf{x} = (x, y, t)$. Then, the dominant spatiotemporal orientation at \mathbf{x} is specified by the eigenvector, $\hat{\mathbf{e}}_s$, corresponding to the smallest eigenvalue of \mathbf{Q} , provided the region contains adequate structure. To interpret $\hat{\mathbf{e}}_s$, in terms of optical flow, \mathbf{v} , the eigenvector must be projected onto the image plane: Let $\hat{\xi}_x$ and $\hat{\xi}_y$ be unit vectors defining the image plane, while $\hat{\mathbf{t}}$ is the unit vector along the temporal direction. Optical flow is then recovered as

$$\mathbf{v} = \left(e_x \hat{\xi}_x + e_y \hat{\xi}_y \right) / e_t,$$

where e_x , e_y and e_t are the projections of $\hat{\mathbf{e}}_s$ on $\hat{\xi}_x$, $\hat{\xi}_y$ and $\hat{\mathbf{t}}$, respectively, i.e.,

$$\begin{aligned} e_x &= \hat{\mathbf{e}}_s \cdot \hat{\xi}_x \\ e_y &= \hat{\mathbf{e}}_s \cdot \hat{\xi}_y \\ e_t &= \hat{\mathbf{e}}_s \cdot \hat{\mathbf{t}}. \end{aligned}$$

Bibliography

- [1] Granlund, G., Knutsson, H.: Signal Processing for Computer Vision. Kluwer (1995)
- [2] Bigun, J.: Vision with Direction. Springer (1998)
- [3] Jenkin, M., Tsotsos, J.K.: Applying temporal constraints to the dynamic stereo problem. Computer Vision, Graphics and Image Processing **33** (1986) 16–32
- [4] Waxman, A.M., Duncan, J.H.: Binocular image flows: Steps toward stereo-motion fusion. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **8** (1986) 715–729
- [5] Weng, J., Cohen, P., Rebibo, N.: Motion and structure estimation from stereo image sequences. IEEE Journal of Robotics and Automation (RA) **8** (1992) 362–382
- [6] Zhang, Z., Faugeras, O.D.: Three-dimensional motion computation and object segmentation in a long sequence of stereo frames. International Journal of Computer Vision (IJCV) **7** (1992) 211–241
- [7] Demirdjian, D., Darrell, T.: Using multiple-hypothesis disparity maps and image velocity for 3D motion estimation. International Journal of Computer Vision (IJCV) **47** (2002) 219–228
- [8] Malassiotis, S., Strintzis, M.G.: Model-based joint motion and structure estimation from stereo images. Computer Vision and Image Understanding (CVIU) **65** (1997) 79–94
- [9] Neumann, J., Aloimonos, Y.: Spatio-temporal stereo using multi-resolution subdivision surfaces. International Journal of Computer Vision (IJCV) **47** (2002) 181–193

- [10] Pons, J.P., Keriven, R., Faugeras, O., Hermosillo, G.: Variational stereo & 3D scene flow estimation with statistical similarity measures. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). (2003) 597–602
- [11] Davis, J., Ramamoorthi, R., Rusinkiewicz, S.: Spacetime stereo: A unifying framework for depth from triangulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* **27** (2005) 296–302
- [12] Zhang, L., Curless, B., Seitz, S.M.: Spacetime stereo: Shape recovery for dynamic scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2003) 367–374
- [13] Gong, M.: Enforcing temporal consistency in real-time stereo estimation. In: Proceedings of the European Conference on Computer Vision (ECCV). (2006) 564–577
- [14] Strecha, C., van Gool, L.: Motion-stereo integration for depth estimation. In: Proceedings of the European Conference on Computer Vision (ECCV). (2002) 170–185
- [15] Huguet, F., Devernay, F.: A variational method for scene flow estimation from stereo sequences. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). (2007) 1–7
- [16] Sudhir, G., Baneerjee, S., Biswas, K.K., Bahl, R.: Cooperative integration of stereopsis and optic flow computation. *JOSA-A* **12** (1995) 2564–2572
- [17] Leung, C., Appleton, B., Lovell, B.C., Sun, C.: An energy minimisation approach to stereo-temporal dense reconstruction. In: Proceedings of the IEEE International Conference on Pattern Recognition (ICPR). (2004) 72–75
- [18] Williams, O., Isard, M., MacCormick, J.: Estimating disparity and occlusions in stereo video sequences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2005) 250–257

- [19] Isard, M., MacCormick, J.: Dense motion and disparity estimation via loopy belief propagation. In: Proceedings of the Asian Conference on Computer Vision (ACCV). Volume 2. (2006) 32–41
- [20] Larsen, E.S., Mordohai, P., Pollefeys, M., Fuchs, H.: Temporally consistent reconstruction from multiple video streams. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). (2007) 1–8
- [21] Hanna, K.J., Okamoto, N.E.: Combining stereo and motion analysis for direct estimation of scene structure. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). (1993) 357–365
- [22] Stein, G.P., Shashua, A.: Direct estimation of motion and extended scene structure from a moving stereo rig. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (1998) 211–218
- [23] Mandelbaum, R., Salgian, G., Sawhney, H.: Correlation-based estimation of ego-motion and structure from motion and stereo. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). (1999) 544–550
- [24] Beauchemin, S., Barron, J.: The computation of optical flow. *ACM Comp. Surv.* **27** (1995) 433–467
- [25] Heeger, D.: A model for the extraction of image flow. *JOSA A* **4** (1997) 1455–1471
- [26] Cannons, K., Wildes, R.P.: Spatiotemporal oriented energy features for visual tracking. In: Proceedings of the Asian Conference on Computer Vision (ACCV). (2007) 532–543
- [27] Derpanis, K.G., Wildes, R.P.: Early spatiotemporal grouping with a distributed oriented energy representation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2009)
- [28] Shechtman, E., Irani, M.: Space-time behavior-based correlation - or - how to tell if two underlying motion fields are similar without computing

- them? *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* **29** (2007) 2045–2056
- [29] Jones, D.G., Malik, J.: A computational framework for determining stereo correspondence from a set of linear spatial filters. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. (1992) 395–410
 - [30] Freeman, W.T., Adelson, E.H.: The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* **13** (1991) 891–906
 - [31] Derpanis, K.G., Gryn, J.: Three-dimensional n-th derivative of Gaussian separable steerable filters. In: *Proceedings of the IEEE International Conference on Image Processing (ICIP)*. Volume 3. (2005) 553–556
 - [32] Shechtman, E., Irani, M.: Space-time behavior based correlation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Volume 1. (2005) 405–412
 - [33] Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: *Proceedings of the International Joint Conference on Artificial Intelligence*. (1981) 674679
 - [34] Baker, S., Matthews, I.: Lucas-Kanade 20 years on: A unifying framework. *International Journal of Computer Vision (IJCV)* **56** (2004) 221–255
 - [35] Sizintsev, M., Wildes, R.P.: Efficient stereo with accurate 3-D boundaries. In: *Proceedings of the British Machine Vision Conference (BMVC)*. (2006) 237–246
 - [36] Kolmogorov, V., Zabih, R.: Computing visual correspondence with occlusions using graph cuts. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. (2001) 508–515
 - [37] Larsen, E.S., Mordohai, P., Pollefeys, M., Fuchs, H.: Temporally consistent reconstruction from multiple video streams using enhanced belief propagation. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. (2007) 1–8

- [38] Point Grey Research: <http://www.ptgrey.com> (2008)
- [39] Scharstein, D., Szeliski, R.: High-accuracy stereo depth maps using structured light. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Volume 1. (2003) 195–202
- [40] Farid, H., Simoncelli, E.P.: Differentiation of discrete multi-dimensional signals. IEEE Transactions on Image Processing **13** (2004) 496–508
- [41] Black, M.J., Anandan, P.: The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. Computer Vision and Image Understanding (CVIU) **61** (1996) 75–104