Detecting Spatiotemporal Structure Boundaries: Beyond Motion Discontinuities

Konstantinos G. Derpanis and Richard P. Wildes

Department of Computer Science and Engineering
York University
Toronto, Ontario, Canada
{kosta,wildes}@cse.yorku.ca

Abstract. The detection of motion boundaries has been and remains a longstanding challenge in computer vision. In this paper, the recovery of motion boundaries is recast in a broader scope, as focus is placed on the more general problem of detecting spacetime structure boundaries, where motion boundaries constitute a special case. This recasting allows uniform consideration of boundaries between a wider class of spacetime patterns than previously considered in the literature, both coherent motion as well as additional dynamic patterns. Examples of dynamic patterns beyond standard motion that are encompassed by the proposed approach include, flicker, transparency and various dynamic textures (e.g., scintillation). Toward this end, a novel representation and method for detecting these boundaries in raw image sequence data are presented. Central to the representation is the description of oriented spacetime structure in a distributed manner. Empirical evaluation of the proposed boundary detector on challenging natural imagery suggests its efficacy.

1 Introduction

The detection of motion boundaries in (temporal) image sequences has been and remains a longstanding challenge in computer vision. The reason for continued interest is due in part to their providing boundary conditions for any process that requires knowledge of the spacetime support of coherent data for recovery of reliable local estimates (e.g., optical flow). In addition, these boundaries provide useful information about the 3D structure of the imaged scene.

Although of obvious importance, motion represents a particular instance of the myriad spatiotemporal patterns encountered in image sequences. Examples of non-motion-related patterns of significance include, unstructured (e.g., "blank wall"), flicker (i.e., pure temporal intensity change), and dynamic texture (e.g., as typically associated with stochastic phenomena, such as windblown vegetation and turbulent water). These types of dynamic patterns have received far less attention than motion in the literature.

The goal of the present work is the development of a unified approach to detecting spacetime boundaries that is broadly applicable to the diverse phenomena encountered in the natural world, including but not limited to motion. It is proposed that the choice of representation is key to meeting this challenge: If the representation cannot adequately distinguish the patterns of interest, then the recovery of boundaries, regardless of the chosen detector, will fail. For present purposes, local 3D, (x, y, t), spacetime orientation will be shown to be of appropriate descriptive power. Measures of spatiotemporal orientation capture the first-order correlation structure of the data irrespective of its origin (e.g., irrespective of its physical cause), even while distinguishing a wide range of patterns of interest (e.g., different motions, as well as the various aforementioned additional dynamic patterns). With visual spacetime represented according to its local orientation structure, boundaries will be extracted via detection of spatiotemporal change in the local orientation structure.

Previous dynamic boundary detection methods can be categorized as either local or global. Local methods restrict analysis to limited neighbourhoods around each point. In contrast, global methods generally attempt to simultaneously estimate a consistent flow field and its discontinuities across the image.

Early efforts focused on the local detection of motion discontinuities in dense optical flow fields through the use of edge operators (e.g., [1]). Alternatively, regions exhibiting a high percentage of unmatched features on a frame-to-frame basis are identified as motion boundaries [2]. Other methods have detected boundaries from the shape of the local template match surface (e.g., [3]). Boundary detection also has been performed using a detector over basis flows for simple events (e.g., motion of occluding edge or bar) [4]. In follow-up work, motion discontinuity regions were captured using a non-linear generative model [5]. Alternatively, hand-labeled motion boundaries have been used to train a discriminative classifier [6]. Further, motion boundary detection has been based on analysis of local distributions of image features (e.g., intensity, colour, flow) [7,8]. Perhaps most closely related to the approach proposed here are methods that detect motion boundaries from the structure of spatiotemporal brightness patterns as captured by local estimates of spatiotemporal orientation [9] or, more generally, oriented bandpass filters [10,11]. Also related are previous efforts using oriented energy measurements for boundary detection in 2D intensity images, e.g., [12,13].

Typically, the focus of global methods has been the recovery of regional flows, with inter-region boundaries made explicit to various degrees [14,15]. The particular formulations developed in these cases are limited to motion boundary detection and not more generally applicable to additional classes of spatiotemporal structure boundaries. Alternatively, global methods have been developed that indicate regions of dynamic texture and their boundaries, e.g., [16]; however, it does not appear that such methods are applicable directly to motion boundaries.

Overall, it appears that no single previous method for spatiotemporal boundary detection is capable of capturing the wide range of juxtaposed spacetime patterns encountered in the real world. Furthermore, the emphasis of most previous work has been on the special case of motion boundaries.

In the light of previous research, the following three major contributions are made. (i) The problem of detecting motion boundaries is recast in terms of the more general problem of identifying spacetime structural boundaries. This recasting allows for capturing, in a unified manner, boundaries between a wide range of important spatiotemporal patterns (unstructured, static, motion, flicker, (pseudo-)transparency, translucency, scintillation). (ii) A new representation is proposed for identifying spatiotemporal boundaries that captures local 3D, (x, y, t), image spacetime orientation structure in a distributed manner. The representation converts structure differences to spatiotemporal contrast; correspondingly, simple contrast detection mechanisms (e.g., local differential operators) can mark boundaries. (iii) The proposed boundary detector's ability to identify boundaries along meaningful structural lines is shown quantitatively and outperforms several extant approaches on a wide range of challenging natural imagery.

2 Technical approach

The proposed approach to spacetime representation and boundary analysis consists of an initial local oriented decomposition of the input video, followed by detecting spacetime structural boundaries across the decomposition. This approach is motivated by the fact that such a decomposition captures significant, meaningful aspects of its temporal variation [11]. As examples: A significant response in a single component of the decomposition is indicative of motion; significant responses in multiple components of the decomposition are indicative of transparency-based superposition; more uniform, yet still significant responses across the entire decomposition are indicative of dynamic texture (e.g., scintillation); lack of response in any component of the decomposition is indicative of unstructured regions (e.g., uniform intensity). Under this representation, coherency of spacetime is defined in terms of consistent patterns across the decomposition, while inconsistencies indicate spacetime structural boundaries. Integration of purely spatial cues (e.g., colour and texture), although of obvious benefit, is beyond the scope of this contribution.

2.1 Spatiotemporal oriented energy representation

The spacetime orientation decomposition is realized using broadly tuned 3D Gaussian second derivative filters, $G_{2_{\hat{\theta}}}(x, y, t)$, and their Hilbert transforms, $H_{2_{\hat{\theta}}}(x, y, t)$, with the unit vector $\hat{\theta}$ capturing the 3D direction of the filter symmetry axis. The responses are pointwise rectified (squared) and summed to yield the following energy measure,

$$E_{\hat{\theta}}(x, y, t) = (G_{2_{\hat{\theta}}} * I)^2 + (H_{2_{\hat{\theta}}} * I)^2, \qquad (1)$$

where $I \equiv I(x, y, t)$ denotes the input imagery and $*$ convolution.

Each oriented energy measure, (1), is confounded with spatial orientation. Consequently, in cases where the spatial structure varies widely about an otherwise coherent dynamic region (e.g., single motion of a surface with varying spatial texture), the responses of the ensemble of oriented energies will reflect this behaviour and thereby support spurious region segregation. To ameliorate this difficulty, the spatial orientation component is discounted by "marginalization" of this attribute, as follows.

In general, a pattern exhibiting a single spacetime orientation (e.g., velocity) manifests itself as a plane through the origin in the frequency domain [17]. Correspondingly, summation across a set of x-y-t-oriented energy measurements consistent with a single frequency domain plane through the origin is indicative of energy along the associated spacetime orientation, independent of purely spatial orientation. Since Gaussian derivative filters of order $N = 2$ are used in the oriented filtering, (1), it is appropriate to consider $N+1=3$ equally spaced directions along each frequency domain plane of interest, as $N+1$ directions are needed to span orientation in a plane with Gaussian derivative filters of order N [13]. Let each plane be parameterized in terms of its unit normal, $\hat{\mathbf{n}}$; a set of equally spaced $N+1$ directions within the plane are given as

$$\hat{\theta}_i = \cos\left(\frac{2\pi i}{N+1}\right)\hat{\theta}_a(\hat{\mathbf{n}}) + \sin\left(\frac{2\pi i}{N+1}\right)\hat{\theta}_b(\hat{\mathbf{n}}), \quad 0 \le i \le N, \tag{2}$$

with

$$\hat{\theta}_a(\hat{\mathbf{n}}) = \hat{\mathbf{n}} \times \hat{\mathbf{e}}_x / \|\hat{\mathbf{n}} \times \hat{\mathbf{e}}_x\| \quad \text{and} \quad \hat{\theta}_b(\hat{\mathbf{n}}) = \hat{\mathbf{n}} \times \hat{\theta}_a(\hat{\mathbf{n}}), \tag{3}$$

where $\hat{\mathbf{e}}_x$ denotes the unit vector along the ω_x-axis[1]. In the case where the spacetime orientation is defined by velocity (u_x, u_y), the normal vector is given by $\hat{\mathbf{n}} = (u_x, u_y, 1)^\top / \|(u_x, u_y, 1)^\top\|$.

Now, energy along a spacetime direction, $\hat{\mathbf{n}}$, with spatial orientation discounted through marginalization, is given by summation across the set of measurements, $E_{\hat{\theta}_i}$,

$$\tilde{E}_{\hat{\mathbf{n}}}(x,y,t) = \sum_{i=0}^{N} E_{\hat{\theta}_i}(x,y,t), \tag{4}$$

with $\hat{\theta}_i$ one of $N+1=3$ specified directions, (2), and each $E_{\hat{\theta}_i}$ calculated via the oriented energy filtering, (1), (cf. [18] where a similar formulation is developed, but only applied to image motion analysis and without inclusion of the $H_{2\theta}$, which provides phase independence). In the present implementation, six different spacetime orientations are made explicit, namely, leftward, rightward, upward and downward motion, static (no motion/orientation orthogonal to the image plane) and flicker/infinite motion (orientation orthogonal to the temporal axis); although, due to the broad tuning of the filters employed, responses arise to a range of orientations about the peak tunings.

Finally, the resulting energies in (4) are confounded by the local contrast of the signal and as a result increase monotonically with contrast. This makes it impossible to determine whether a high response for a particular spacetime orientation is indicative of its presence or is instead a low match that yields a high response due to significant contrast in the signal. To arrive at a purer measure of spacetime orientation, the energy measures are normalized by the sum of consort planar energy responses at each point,

$$\hat{E}_{\hat{\mathbf{n}}_i}(x,y,t) = \tilde{E}_{\hat{\mathbf{n}}_i}(x,y,t) / \left(\sum_{j=1}^{M} \tilde{E}_{\hat{\mathbf{n}}_j}(x,y,t) + \epsilon\right), \tag{5}$$

where M denotes the number of spacetime orientations considered, and ϵ a constant introduced as a noise floor and to avoid instabilities at points where the overall energy is small. Conceptually, (1) - (5) can be thought of as taking an image sequence, $I(x,y,t)$, and carving its (local) power spectrum into a set of planes, with each plane corresponding to a particular spacetime orientation, to provide a relative indication of the presence of structure along each plane.

[1] Depending on the spacetime orientation sought, $\hat{\mathbf{e}}_x$ can be replaced with another axis to avoid the case of an undefined normal vector.

The constructed representation enjoys a number of attributes that are worth emphasizing. (i) Owing to the bandpass nature of the Gaussian derivative filters (1), the representation is invariant to additive photometric bias. (ii) Owing to the normalization (5), the representation is invariant to absolute contrast in the input signal. (iii) Owing to the marginalization (4), the representation is invariant to changes in appearance manifest as spatial orientation variation. Overall, these three invariances result in robust boundary detection that is invariant to pattern changes that do not correspond to dynamic pattern variation, even while making explicit local orientation structure that arises with temporal variation (motion, flicker, scintillation, etc.). (iv) The representation is efficiently realized via linear (separable convolution, pointwise addition) and pointwise non-linear (squaring, division) operations [19].

2.2 Anisotropic smoothing

Prior to attempting to mark loci of significant spatiotemporal boundaries in the oriented energy decomposition, it is appropriate to smooth the derived representation to suppress noise. For this purpose, an anisotropic smoothing is performed as it serves to attenuate noise while enhancing structural boundaries. In the current implementation, mean-shift is employed as the anisotropic smoothing operation [20]. To promote spatiotemporal coherence at the smoothing stage, the orientation feature-space is augmented with positional information in the form of spacetime coordinates, (x, y, t). Putting the above features together yields a 9D feature vector (six oriented energies plus three for spacetime location), per image point.

Conceptually, mean-shift regards the feature-space as an empirical distribution. Each feature-point is associated with a mode (local maximum) of the distribution and thereby all points associated with a particular mode share a common feature value. In its simplest formulation (i.e., based on the Epanechnikov kernel), the mean-shift property can be written as (see [20], for details)

$$\widehat{\nabla} f(\mathbf{x}_c) \propto \left(\underset{\mathbf{x}_i \in \mathbf{S}_{h,\mathbf{x}_c}}{\text{mean}} \{\mathbf{x}_i\} - \mathbf{x}_c \right), \tag{6}$$

where $f(\mathbf{x})$ denotes the underlying probability density function of a n-dimensional space, $\mathbf{x}$, $\{\mathbf{x}_i\}$ the given set of samples, and $\mathbf{S}_{h,\mathbf{x}_c}$ a n-dimensional hyper-ball with radius h (the so-called kernel density bandwidth) centered at $\mathbf{x}_c$. Repeated application of (6) converges to a local mode of the distribution. In the present case, modes arise as particular values across 9D spatiotemporal feature vectors, $\mathbf{x}$. The final smoothed energy representation is realized by assigning the converged oriented energy portion of the feature vectors to their respective initial spacetime positions.

2.3 Spatiotemporal structure boundaries

In essence, the oriented energy representation converts spacetime structure differences to intensity differences across its decomposition. Correspondingly, boundaries simply correspond to image loci exhibiting significant spatiotemporal contrast in the representation. Figure 1 illustrates this point. In the orientation decomposition, it is seen that the foreground tree yields relatively large and small intensities in the "static" and "rightward" components (resp.); whereas, the moving background yields the opposite behaviour. Therefore, spatiotemporal change (i.e., contrast) in the decomposition is indicative of the boundary between the tree and background. More generally, the orientation decomposition is a multivalued image, with spatiotemporal contrast indicative of spacetime boundaries in the underlying data. Here, it is interesting to note the difference in the behaviour of flow estimates and the proposed distributed representation across boundaries. In the former, the results are unpredictable due to a total failure of its intrinsic assumptions (e.g., brightness conservation). In the latter, due to the considerable overlap in spacetime and orientation tuning of the filters, the representation changes *smoothly* across structure boundaries reflecting the shift of energies among channels.

To capture the spatiotemporal contrast in the (smoothed) oriented energy representation, (5), a generalized gradient formulation is employed, as it captures change in a uniform manner across the multiple components of the decomposition. Let $\hat{E}_k$ be the kth band of the oriented energy representation, (5), and $\xi_i = x, y, t$ for $i = 1, 2, 3$, resp., define the directions along which partial derivatives are taken, then the generalized gradient is a 3×3 matrix $\mathbf{S}$ where

$$\mathbf{S}_{ij} \equiv \sum_{k=1}^{n} (\partial \hat{E}_{\hat{\mathbf{n}}_k} / \partial \xi_i)(\partial \hat{E}_{\hat{\mathbf{n}}_k} / \partial \xi_j). \tag{7}$$

Notice that $\mathbf{S}$ amounts to the summation of the more standard *structure/gradient tensor* [21] of each energy band[2]. The eigenvector of (7) associated with the greatest eigenvalue, λ_1, denoted $\mathbf{e}_1$, points in the direction of greatest change in the feature-space. For multivalued images (i.e., $n > 1$), a boundary is not indicated simply by a large value for λ_1; instead, it must be large relative to the other eigenvalues of $\mathbf{S}$ [22]. Correspondingly, a normalized measure of spacetime structure boundary salience is employed in the present context

$$\text{boundary}_{\text{salience}} = (\lambda_1 - \lambda_2)/(\lambda_1 + \lambda_2 + \phi), \tag{8}$$

where $\lambda_1 > \lambda_2$ denote the two largest eigenvalues of $\mathbf{S}$ and ϕ is a constant introduced as a noise floor. High values of the boundary salience measure, (8), (i.e, values close to one), are indicative of the presence of a spacetime structure boundary. Boundary salience for the example in Fig. 1 is shown in its rightmost panel. Next, similar to the non-maximum suppression principle used in intensity-based edge detection [24], a candidate boundary point is defined as a point that achieves a maximum in boundary salience, (8), in the direction of the eigenvector $\mathbf{e}_1$, as follow,

$$\begin{cases} \dfrac{\partial \text{boundary}_{\text{saliency}}}{\partial \mathbf{e}_1} = 0 \\ \dfrac{\partial^2 \text{boundary}_{\text{saliency}}}{\partial \mathbf{e}_1^2} < 0 \end{cases}. \tag{9}$$

Finally, candidate loci having a saliency value greater than a certain threshold, τ, are marked as boundary points.

[2] Other adaptations of the generalized gradient to multiband image boundary detection include application to colour [22] and spatial texture [23].

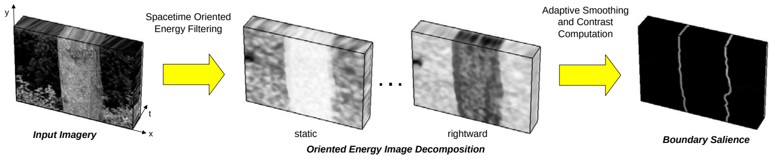

Fig. 1. Oriented energy decomposition maps structural differences to intensity differences. (left) Input image sequence of a foreground tree tracked (stabilized) by a moving camera with background in relative motion. (middle) Oriented energy decomposition of input shows marked differences in intensity corresponding to dynamic pattern differences of foreground vs. background. (right) Boundaries marked according to spatiotemporal contrast across the energy decomposition.

2.4 Algorithm

To recapitulate, the proposed approach can be given in algorithmic terms as follows.

```
Input: Greyscale image sequence
Input parameter: Boundary detection threshold, $\tau$
Output: Binary image sequence marking spatiotemporal structure boundaries

Step 1: Compute spacetime oriented energy representation (Section 2.1)
  1. Initialize 3D $G_2/H_2$ steerable basis.
  2. Compute normalized spacetime oriented energy measure, Eqs. (1)-(5).
Step 2: Anisotropic smoothing: Mean-shift (Section 2.2)
  1. Augment each normalized spacetime oriented energy measure, (5), with its
     spacetime coordinate $(x, y, t)$.
  2. Apply mean-shift smoothing iterations, (6).
  3. Replace each energy measure in (5) with the final converged energy measure.
Step 3: Compute spatiotemporal structure boundary salience (Section 2.3)
  for each spacetime point
  1. Construct generalized gradient, (7), from (smoothed) oriented energy represen-
     tation, (5).
  2. Compute the eigenvector/eigenvalues of the generalized gradient, (7).
  3. Compute boundary salience, (8).
Step 4: Non-maximum suppression (Section 2.3)
  for each spacetime point
  1. Apply non-maximum suppression, (9).
  2. Retain candidate boundaries that have a saliency value, (8), greater than $\tau$.
```

3 Empirical evaluation

In evaluation, parameter settings for the proposed detector are as follows. The ϵ bias for contrast normalization, (5), empirically has been set to $\approx 1\%$ of the maximum expected response. The noise floor, ϕ, for boundary salience, (8), empirically has been set to $\phi = 0.01$. Mean-shift (anisotropic) smoothing includes three bandwidth parameters, h_{space}, h_{time} and h_{range}, which determine the resolution of detail along the spatial, temporal and range (here, spacetime orientation) dimensions, resp. Unless otherwise stated, the mean-shift bandwidths are set to: $h_{\text{space}} = 32$, $h_{\text{time}} = 10$, and $h_{\text{range}} = 0.12$.

Figure 4 shows a set of challenging natural image sequences containing a broad range of juxtaposed spacetime structures, including but not restricted to motion, and their boundary detection results (see caption for description of inputs). The challenging aspects of this data set include, regions that are unstructured, exhibit significant temporal aliasing due to fast motion, contain superimposed motion (transparency) and non-motion structure (e.g., flicker and scintillation). Coherent motion boundaries constitute a small fraction of the boundaries present in the data. Alternative available data sets are limited by their restricted focus on motion boundaries at the expense of more general spacetime structural boundaries [8]. The sequences presented here, consisting of juxtaposed natural and man-made structures, were obtained from a variety of sources: a Canon HF10 camcorder, the BBC documentary "Planet Earth" and the "BBC Motion Gallery" online video repository. Each sequence spans 10 frames. For each example, frame-by-frame hand-labeled ground truth was established. The identified boundaries in Fig. 4 provide compelling qualitative evidence that the proposed detector performs well on image sequences containing a wide variety of spacetime structures. This data set is available at `www.cse.yorku.ca/vision/research/spacetime-grouping`.

To quantify performance, results of the proposed detector are compared with the hand-labeled ground truth as well as alternative approaches. In particular, mean precision/recall scores [25] were calculated across all image sequences shown in Fig. 4 and are shown as tuning curves in Fig. 2 as detection parameters are varied. Here, over-partitioning is characterized in the curves by high recall but low precision, and the converse holds for under-partitioned image sequences.

The left panel of Fig. 2 shows several different curves for the proposed method, with each curve corresponding to a different value of the smoothing parameter, h_{range}; all curves are swept as the detection threshold varies from $0 - 1$. Matching between ground truth and identified boundary points was carried out using a distance threshold of eight, which is reasonable given that the support of the various compared detectors span approximately eight pixels. The consistently high recall indicates that ground truth boundaries are accurately marked. At the same time, a relatively high precision is attained, which indicates false boundaries are not prevalent. Further, the approach is seen to be stable with respect to variation of the smoothing parameter.

The right panel of Fig. 2 compares the best curve of the proposed approach, $h_{range} = 0.14$, with two alternative methods: (1) edge detection on dense optical flow fields [1] (implemented as a 3D Canny edge operator [24] applied to flow recovered using Lucas-Kanade [26]) and (2) the rank-based method that analyzes the gradient structure tensor over a neighbourhood [9]. These methods are selected for comparison as they are local (like the proposed method) and edge-detection in flow fields is a long standing approach, while the rank-based analysis is a recent proposal that has shown strong results for certain boundary types. Tuning curves were swept for the flow- and rank-based de-

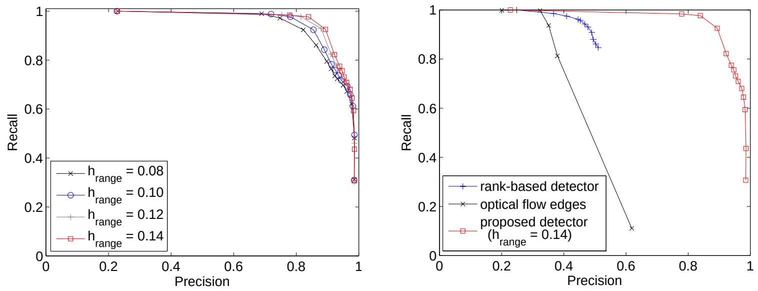

Fig. 2. Precision/recall curves. (left) Precision/recall of the proposed detector, each curve corresponds to a different setting of the range bandwidth used for smoothing. (right) Comparison of precision/recall with the proposed (optimal curve in (left)), flow- and rank-based detectors.

tectors by varying their detection thresholds from $0 - 10$ and $0 - 1$, resp. Curves for all three methods have the expected shape; however, flow- and rank-based are translated along the precision axis, which indicates significant over-partitioning relative to the proposed approach. Along these lines, rank-based outperforms flow, but is still noticeably worse than the proposed method.

To scrutinize the results in Fig. 2 further, Fig. 3 shows a comparison of the various boundary detectors on selected examples from Fig. 4 (c), (f) and (i). Also to compare against global methods, results from a recent level-set-based approach are shown [15]. Note that the global method must be supplied with a priori knowledge of the number of regions and hand initialization of its boundary[3]. For the motion parallax example, all of the alternative methods yield reasonable results. This is to be expected, as they are designed for motion boundaries. In the other two examples, the flow and rank methods yield spurious boundaries in the transparency and scintillation regions. This shortcoming arises from the inability of these methods to recover coherent measurements in non-coherent motion regions, as the assumption that coherency is well characterized by a single smoothly varying flow is violated. These spurious boundaries are the source of the low precision yet high recall rates indicated for the flow- and rank-based detectors in Fig. 2. For the transparency case using level-sets, the part of the initial contour that is outside the moving target evolves correctly; however, the part that started inside the moving region converges incorrectly, as the target interior does not conform to the method's assumption of a single smooth flow. In the scintillation case, the level-set collapses to a single region. Here, the failure is due to the relative lack of spatial structure in the ship interior, which allows the approach to fit a flow across the ship that is consistent with whatever flow it (erroneously) recovers for the scintillating water. The relative lack of structure in the ship interior also accounts for the apparent difference in performance of the flow and rank methods in such regions: Flow recovers highly variable

[3] Due to the dependence of the extracted boundaries on hand contour initialization, number of regions and various scaling parameters in the level-set approach, it is not customary to sweep precision/recall curves for level-sets; hence, only qualitative comparisons are provided here.

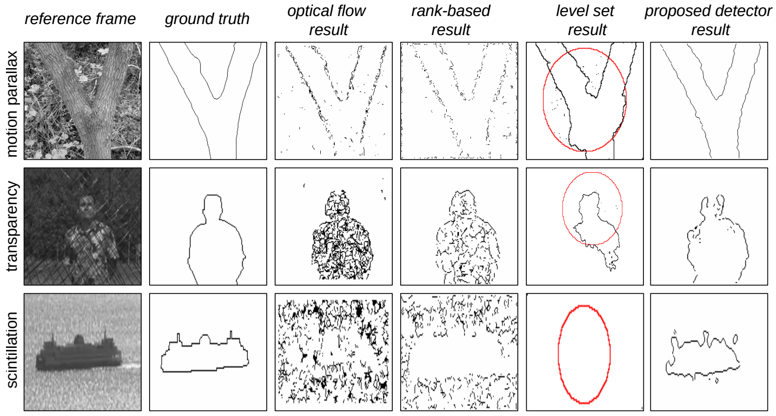

Fig. 3. Comparison of results for flow, rank, level-set and proposed approaches to boundary detection applied to Fig. 4 (c), (f) and (i). For level sets, red and black curves show hand initialized and converged result, resp. For the scintillation example the level set collapses to yield a single region.

vector fields that are interpreted as boundaries; whereas, unstructured regions are rank consistent and thus do not yield spurious boundaries. In contrast to the alternatives, the proposed detector naturally handles all three cases highlighted in Fig. 3.

4 Discussion and summary

Most previous methods for spatiotemporal boundary detection are concerned with borders between regions of contrasting optical flow. Others are focused on dynamic textures. Improvements to these various methods might be realized via introduction of thresholds (e.g., confidence measures), multi-scale analyses (e.g., pyramid schemes for accommodating rapid motion), contour completion (cf. [9]), a more sophisticated flow estimator than considered here, etc. These approaches, however, fundamentally are limited by their underlying assumptions regarding the classes of visual phenomena that are to be encountered, which in turn limit their applicability to detecting a very circumscribed class of boundaries (e.g., motion). In comparison, it has been demonstrated that the proposed approach can naturally deal with the wide variety of real-world scenarios presented.

In summary, this paper has presented a unified approach to representing and detecting boundaries between a wide range of juxtaposed spacetime patterns (unstructured, static, motion, flicker, (pseudo-)transparency, translucency, scintillation). The approach is based on a distributed characterization of visual spacetime in terms of 3D, (x, y, t), spatiotemporal orientation, followed by application of a spatiotemporal differential operator (generalized gradient) to mark boundaries. Empirical evaluation on a wide variety of imagery demonstrates the proposed detector's ability to delineate boundaries between coherently structured regions.

Acknowledgments. This research was supported in part by NSERC and Precarn. F. Estrada provided software for computing precision/recall curves and helpful discussion.

References

1. Thompson, W., Mutch, K., Berzins, V.: Dynamic occlusion analysis in optical flow fields. PAMI **7** (1985) 374–383
2. Mutch, K., Thompson, W.: Analysis of accretion and deletion at boundaries in dynamic scenes. PAMI **7** (1985) 133–138
3. Anandan, P.: Computing dense fields displacement with confidence measures in scenes containing occlusion. In: DARPA IUW. (1984) 236–246
4. Fleet, D., Black, M., Jepson, A.: Motion feature detection using steerable flow fields. In: CVPR. (1998) 274–281
5. Black, M., Fleet, D.: Probabilistic detection and tracking of motion boundaries. IJCV **38** (2000) 231–245
6. Apostoloff, N., Fitzgibbon, A.: Learning spatiotemporal T-junctions for occlusion detection. In: CVPR. (2005) II: 553–559
7. Spoerri, A., Ullman, S.: The early detection of motion boundaries. In: ICCV. (1987) 209–218
8. Stein, A., Hebert, M.: Local detection of occlusion boundaries in video. IVC **27** (2009) 514–522
9. Feldman, D., Weinshall, D.: Motion segmentation and depth ordering using an occlusion detector. PAMI **30** (2008) 1171–1185
10. Niyogi, S.: Detecting kinetic occlusion. In: ICCV. (1995) 1044–1049
11. Wildes, R., Bergen, J.: Qualitative spatiotemporal analysis using an oriented energy representation. In: ECCV. (2000) II: 768–784
12. Morrone, M., Owens, R.: Feature detection from local energy. PRL **6** (1987) 303–313
13. Freeman, W., Adelson, E.: The design and use of steerable filters. PAMI **13** (1991) 891–906
14. Heitz, F., Bouthemy, P.: Multimodal estimation of discontinuous optical flow using Markov random fields. PAMI **15** (1993) 1217–1232
15. Cremers, D., Soatto, S.: Motion competition: A variational approach to piecewise parametric motion segmentation. IJCV **62** (2005) 249–265
16. Doretto, G., Cremers, D., Favaro, P., Soatto, S.: Dynamic texture segmentation. In: ICCV. (2003) 1236–1242
17. Watson, A., Ahumada, Jr., A.: A look at motion in the frequency domain. In: Motion Workshop. (1983) 1–10
18. Simoncelli, E.: Distributed Analysis and Representation of Visual Motion. PhD thesis, MIT (1993)
19. Derpanis, K., Gryn, J.: Three-dimensional nth derivative of Gaussian separable steerable filters. In: ICIP. (2005) III: 553–556
20. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. PAMI **24** (2002) 603–619
21. Jähne, B.: Digital Image Processing, sixth edition. Springer-Verlag, Berlin (2005)
22. Sapiro, G., Ringach, D.: Anisotropic diffusion of multivalued images with applications to color filtering. T-IP **5** (1996) 1582–1586
23. Rubner, Y., Tomasi, C.: Coalescing texture descriptors. In: ARPA IUW. (1996) 927–935
24. Canny, J.: A computational approach to edge detection. PAMI **8** (1986) 679–698
25. Estrada, F., Jepson, A.: Benchmarking image segmentation algorithms (to appear). IJCV (2009)
26. Lucas, B., Kanade, T.: An iterative registration technique with an application to stereo vision. In: IJCAI. (1981) 674–679

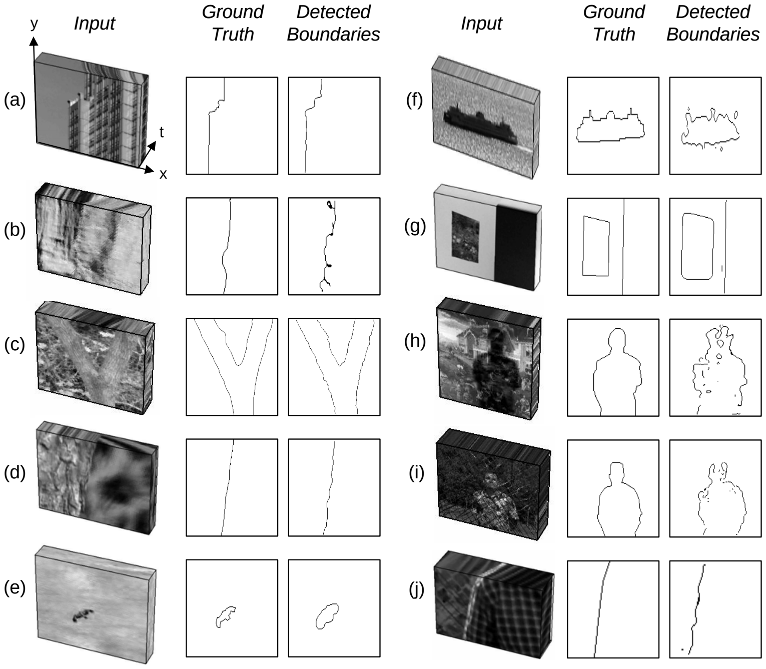

Fig. 4. Boundary detection results on a diverse and challenging set of natural imagery. In each example, the input sequence, a frame from the human-labeled ground truth and the boundary detection result, resp. are given. (a) A panning sequence consisting of a clear sky (i.e., unstructured) and a building (source: HF10). (b) Motion parallax sequence consisting of two mountain faces, where the foreground surface moves rapidly revealing a slower moving surface (source: "Planet Earth"). (c) Tree in foreground being coarsely stabilized by moving camera operator with resulting background motion (source: HF10). The background consisting of the ground plane is not fronto-parallel with respect to the camera, as a result the motion varies across the surface. (d) A leopard rapidly moving leftward behind a static tree (source: "Planet Earth"). (e) A flying bird crudely tracked by the camera operator to yield a slow moving target and a rapidly moving background (source: "Planet Earth"). (f) A ship moving over a scintillating water surface (source: "BBC Motion Gallery"). (g) A painting hanging on an unstructured wall with a light flickering in an adjacent hallway (source: HF10). (h) A translucency sequence realized by projecting (using an LCD projector) a walking person over a static painting (source: HF10). (i) A pseudo-transparency sequence consisting of a person walking behind a fence (source: HF10). (j) A juxtaposed motion and pseudo-transparency sequence consisting of two people moving rightward, one moving in front of a fence while the second is moving behind it (source: HF10). To view these videos, see supplemental material.