Early Spatiotemporal Grouping with a Distributed Oriented Energy Representation

Konstantinos G. Derpanis and Richard P. Wildes
Department of Computer Science and Engineering
York University
Toronto, Ontario, Canada
{kosta,wildes}@cse.yorku.ca

Abstract

Spatiotemporal data is associated with vast amounts of raw samples. Given the limited computational resources typically available, an initial organization of this data supporting semantically meaningful lines of inquiry would facilitate efficient processing. In this paper, a new representation for grouping raw image data into a set of coherent spacetime regions is proposed. Unique in this proposal is that coherency is related to a richer description of local spacetime structure than generally considered. In particular, the representation describes the presence of particular oriented spacetime structures in a distributed manner. A key advantage of this representation is its ability to signal the presence of multiple oriented structures at a given spacetime location. More generally, the abstraction allows for the description and grouping of motion and non-motion-related patterns in a uniform manner. Empirical evaluation of the grouping method on synthetic and challenging natural imagery suggests its efficacy.

1. Introduction

1.1. Motivation

Processing of temporal image sequences can be facilitated by an early grouping of the visual data into coherent spacetime regions. Operations that can benefit from such an organization include target recognition and tracking, parametric motion analysis, video indexing, compression and coding. For all of these cases, delineated groupings serve to define support regions that can be processed as wholes for compact and efficient characterization.

A key challenge to spatiotemporal grouping arises from the wide range of naturally occurring phenomena that must be encompassed. The left panel of Fig. 1 shows a natural scene containing several phenomena that should be grouped into distinct regions. The various depicted areas can be characterized as single motion, pseudo-transparency (multiple superimposed motions), scintillation (dynamic/stochastic texture), static (no motion) and unstructured (lacking in enough spatiotemporal contrast to characterize). More generally, additional phenomena are likely to be encountered, including variations on transparency (e.g., translucency) and rapid brightness change (flicker). Critically, motion encompasses only a subset of the spatiotemporal patterns that must be captured in grouping that is widely applicable to imagery of the natural world. Extant approaches deal with such diversity on a case-by-case basis, with motion predominant. Indeed, it appears that no single previous approach to spatiotemporal grouping can be applied with success to a wide range of natural phenomena.

The goal of the present work is the development of a unified approach to spatiotemporal grouping that is broadly applicable to the diverse phenomena encountered in the natural world. It is proposed that the choice of representation is key to meeting this challenge: If the representation cannot adequately characterize and distinguish the patterns of interest, then no subsequent grouping algorithm will make the appropriate delineations. For present purposes, local spatiotemporal orientation is of fundamental descriptive power, as it captures the first-order correlation structure of the data irrespective of its origin (i.e., irrespective of the underlying visual phenomena). Correspondingly, visual spacetime will be represented according to its local 3D, (x, y, t), orientation structure. In particular, each point of spacetime will be associated with a distribution of measurements that indicates the relative presence of a particular set of spatiotemporal orientations. The middle panel of Fig. 1 shows a decomposition of the working example according to its local orientation structure.

Given the orientation-based representation, distinct regions are defined as groups of (x, y, t) pixels coalesced according to similarities of their orientation distributions. The right panel of Fig. 1 shows the final grouping achieved by the proposed approach for the working example. The results

978-1-4244-3991-1/09/$25.00 ©2009 IEEE

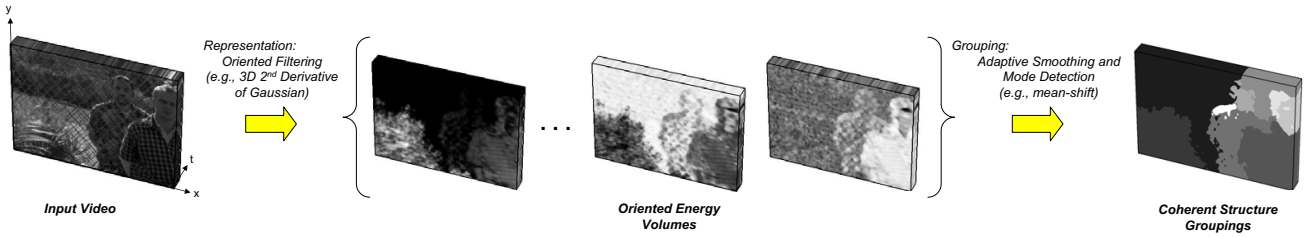

Figure 1. Overview of approach to spatiotemporal analysis. (left) An image sequence serves as input. (middle) Application of energy filters decomposes the input into a distributed representation according to 3D, (x, y, t), spatiotemporal orientation. The example filter outputs are selective for (left-to-right) flicker (purely temporal variation in image brightness), static (no motion) and rightward (motion) spacetime structure. (right) A grouping process operates across the representation to coalesce regions of common spatiotemporal orientation distributions. In the depicted example, input video (left) captures a complex scene: In the central region, a person moves rightward behind a chain-link fence to yield (pseudo-) transparency; a second person to the right of the first person moves rightward in front of the fence to yield a single dominant motion; at the available resolution of the image data, the facial regions have little spatial variation and thereby yield unstructured regions; a windblown plant occupies the left area to yield a scintillating pattern; the remaining region made up of the fence and varying background yields a static pattern. The output region grouping (right) accurately indicates the five major structural regions (transparency, single motion, scintillation, static and unstructured) as five different grey-levels. Depicted are actual results recovered by the approach described in Sec. 2. All data and results shown in this paper are available at: `http://www.cse.yorku.ca/vision/research/spacetime-grouping`.

properly delineate the various groupings even given their diverse nature. Significantly, subsequent analysis can build on the proposed representation, e.g., both dynamic pattern recognition and image motion analysis can exploit a distributed oriented image decomposition [22, 26].

1.2. Related work

The present work touches on two main lines of image sequence analysis: representation of spacetime data and grouping analysis. In terms of representation, most previous efforts centre on the notion of *visual motion* or *temporal/dynamic textures*, with the former garnering the most attention. For visual motion, the majority of work has focused on estimating *optical flow*, which is assumed to be in close correspondence with the *visual motion field* [1]. Previous work also has considered the categorization of the local spacetime orientation structure by way of an eigenvalue analysis of the local orientation tensor [16, 19]. Temporal/dynamic texture related work is concerned with representing and categorizing stochastic dynamic phenomena [20, 12, 4], e.g., turbulent water and windblown vegetation. Each of these strands of research covers a largely disjoint portion of the space of visual spacetime phenomena.

Spatiotemporal oriented energy filters serve in defining the representation employed in the current work. Previous efforts have used similar operators in the analysis of image sequences with application to optical flow estimation [18, 22, 16], motion layer separation [8], activity recognition [5, 11], pattern categorization [26], tracking [3] and spacetime stereo [23]. Significantly, it appears that no previous work has used the filter outputs to drive spatiotemporal grouping, as shown here.

Grouping entails associating together tokens that respect a measure of coherency. Generally, spacetime grouping methods fall into one of two basic computational paradigms: *sequential* and *multi-dimensional* methods. Sequential approaches interleave spatial and temporal grouping processes [24]. These methods are prone to propagating errors across the grouping stages. Multidimensional approaches treat the image sequence and its extracted local attributes as a higher-dimensional feature-space, where grouping is performed. Examples of these methods include, variational [7, 2], voting [21], graph partitioning [14], statistical parametric [17] and non-parametric (e.g., *mean-shift* [9]). Common to the cited approaches is the neglect of the rich underlying local structure of the image data: Grouping is based on overly restrictive measurements, e.g., of colour and optical flow. In contrast, the proposed approach to grouping leverages the more descriptive spatiotemporal orientation structure of the data. As a grouping algorithm mean-shift is employed; however, any of the above spatiotemporal grouping schemes could be used, given the emphasis is on the choice of representation.

1.3. Contributions

In the light of previous research, the major contributions of the present work are as follows. (i) A framework is proposed that yields the first unified approach to representing and grouping a wide range of juxtaposed spacetime patterns (motion, static, flicker, (pseudo-)transparency, translucency, scintillation/temporal texture, unstructured). Previous analyses of spacetime patterns operate on a case-by-case basis, with image motion predominant. (ii) A particular spatiotemporal filtering formulation is developed for measuring spatiotemporal oriented energy and is used as input to a standard grouping mechanism (mean-shift). While spacetime filters have been used before for analyzing image sequences, they have not been applied to the delineation of coherently structured spacetime regions. (iii) The approach's ability to group image data in terms of meaningful

spatiotemporal structure is demonstrated both qualitatively and quantitatively on a wide range of synthetic and challenging natural image sequences.

2. Technical approach

The proposed approach to spatiotemporal grouping consists of an initial local, oriented decomposition of the input datastream, followed by spatiotemporal aggregation across the decomposition that reveals intra-region similarity. The major assumption used in the grouping analysis is smoothness of spacetime structure surfaces in a higher-dimensional feature-space for each coherent region (cf. [21]). Invoking this assumption avoids the burden of explicitly modeling the set of spatiotemporal structures that may be encountered.

2.1. Distributed spatiotemporal orientation

The local spacetime orientation of a visual pattern captures significant, meaningful aspects of its temporal variation (e.g., static vs. moving vs. unstructured) [26]; therefore, a spatiotemporal oriented decomposition of an input pattern is an appropriate basis for spacetime grouping. Under this representation, coherency of spacetime is defined in terms of consistent patterns across the decomposition. A contiguous region of spacetime might be grouped on the basis of it giving rise to a significant response in only one component of the decomposition, corresponding to a particular single orientation (e.g., motion); another region might be grouped because it gives rise to significant responses in multiple components of the decomposition, corresponding to transparency-based superposition; yet another region might be grouped as significant, approximately equal magnitude responses arise across all bands of the decomposition, corresponding to scintillation. Sill other regions might arise as they are distinguished by their lack of local structure. Indeed, Figs. 1 and 2 show such a five-way grouping.

The desired 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 [25]. 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 [15]. 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

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

with

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

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, $E_{\hat{\theta}_i}$, as

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

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. [22] 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 energy measurements 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 indeed 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

[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.

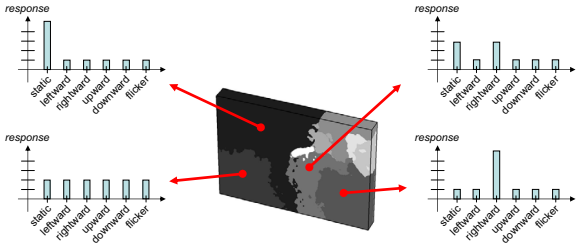

Figure 2. Distributed representation of visual spacetime. The grouping result (centre) is taken from Fig. 1. Each image location carries a distribution representing the degree that the local spacetime structure is consistent with a particular 3D spatiotemporal orientation. The central region of the input, corresponding to transparency, gives rise to two peaks (moving object and static pseudo-transparent foreground); whereas the rightmost region of the input, corresponding to a single moving object, gives rise to a single peak. The leftmost bottom region, corresponding to scintillation, gives rise to a uniform distribution. The facial regions, being devoid of discernable pattern structure at the available resolution, give rise to near zero responses (not shown). The region not attributed to the people or plant, corresponding to static, yields a single peak.

by the sum of consort planar energy responses at each point,

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

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 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.

The constructed representation enjoys a number of attributes that are worth emphasizing. First, owing to the bandpass nature of the Gaussian derivative filters (1), the representation is invariant to additive photometric bias. Second, owing to the normalization (5), the representation is invariant to absolute contrast in the input signal. Third, owing to the marginalization (4), the representation is invariant to changes in appearance manifest as spatial orientation variation. Overall, these three invariances allow grouping abstractions to be robust 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.). Fourth, the representation is efficiently realized via linear (separable convolution, pointwise addition) and pointwise non-linear (squaring, division) operations; thus, efficient computations are realized [10]. Overall, each of the normalized oriented energies can be viewed as expressing the evidence for the presence of a particular, spacetime oriented structure, see, Fig. 2.

2.2. Grouping

To group the oriented energy feature vectors into coherent regions of structure, mean-shift clustering [6] is employed. Note, however, that the focus of the current work is the underlying representation (i.e., distributed measurements of spacetime orientation) rather than a particular grouping mechanism. For the purpose of promoting spacetime coherency within recovered clusters, positional information is included in the feature vector as coordinates in space (x,y) and time t. Putting the above features together yields a 9D feature vector (six oriented energies plus three for spacetime location), per image point. Implicitly this grouping strategy enforces spatial and temporal smoothness simultaneously in the 9D space. This strategy avoids additional assumptions on the camera model, parametric motion model (e.g., translation, affine, etc.) or the number of distinct coherent spacetime structure regions.

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 are grouped together. In its simplest formulation (i.e., based on the Epanechnikov kernel), the mean-shift property can be written as (see [6], for details)

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

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 the 9D spatiotemporal feature vectors and all (x,y,t) pixels contributing to a mode are grouped.

3. Empirical evaluation

A variety of grouping experiments have been performed on a set of synthetic and real world image sequences. Algorithm parameter settings are as follows. The ϵ bias for contrast normalization, (5), empirically has been set to $\approx 1\%$ of the maximum expected response. Mean-shift clustering has four input parameters: the bandwidths, h_{space}, h_{time} and h_{range}, which determine the resolution of mode detection along the spatial, temporal and range (here, spacetime orientation) dimensions, resp., and the merging threshold, τ, which determines the distance between points that are grouped via transitive closure in the 9D feature space. Unless otherwise stated, the mean-shift bandwidth parameters for the space, time and range dimensions are set to $h_{\text{space}} = 32$, $h_{\text{time}} = 10$, and $h_{\text{range}} = 0.12$, resp. Groupings containing less than 40 pixels/frame are considered outliers and removed via merging.

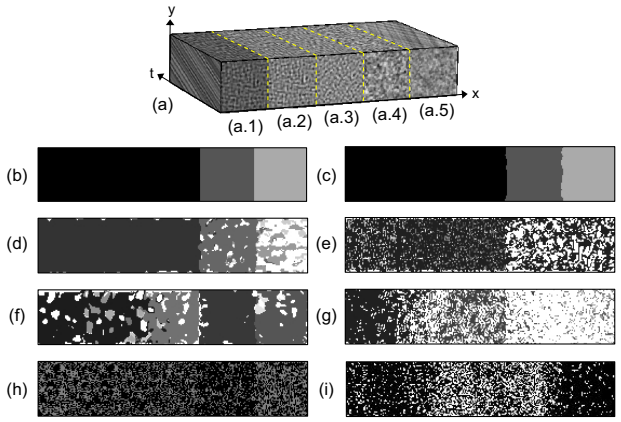

Figure 3. Grouping with various levels of data abstraction. (a) Synthetic input image sequence. (a.1) Bandpass white noise with spatial horizontal/vertical structure enhanced *moving upward*, (a.2) same pattern as previous with contrast increased by 25%, (a.3) bandpass white noise with spatial diagonal structures enhanced moving upward, (a.4) spacetime pink noise (*scintillation*) and (a.5) superimposed pink noise patterns moving rightward/leftward (*transparency*). All motion-related structures move with a speed of 1 pixel/frame. (b) Ideal grouping result with shadings denoting coherent groupings. (c)-(h) Grouping results based on: (c) proposed representation, (d) optical flow, (e) normal flow (f) normalized oriented energy (5) without appearance marginalization (4), (g) spacetime gradient, (h) thresholded frame-to-frame intensity difference, and (i) raw pixel-wise intensity; white denotes outlier components. For each representation, the mean-shift parameters were empirically optimized.

3.1. Representation comparisons

Figure 3 shows a comparison of mean-shift grouping results based on several common spacetime representations in the literature and the proposed distributed representation. The input contains three distinct spacetime structures consisting of upward motion, a scintillating pattern in the form of pink noise and transparency in the form of left/right superimposed motion. The upward moving region contains three subregions that differ only in spatial appearance and therefore should be grouped together from a spatiotemporal perspective due to their common dynamic structure (i.e., upward motion).

Representations that do not adequately discount pure spatial appearance (normalized oriented energy (5) without appearance marginalization (4), spatiotemporal gradient, normal flow and raw pixel-wise intensity) correspondingly fail to support grouping of the coherently moving region as such. Successive frame differencing fails in making the pattern distinctions sought because it is limited to making binary distinctions between stationary and non-stationary patterns. All of these approaches also fail to group successfully the transparency and scintillation cases. While optical flow (estimated as in [16]) proves capable of supporting correct grouping of the coherently upward moving regions, it is incapable of dealing with scintillating and transparency patterns due to violations of the underlying assumption of brightness conservation. Among these abstractions, only the proposed distributed representation successfully supports the partition of the input into its three constituent structures, coherent motion, scintillation and superimposed motion.

3.2. Quantitative results on natural imagery

Figure 4 shows a set of challenging natural image sequences containing a broad range of spacetime structures, including but not restricted to motion, and their grouping 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 and non-motion structure (e.g., transparency and scintillation). These sequences, 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.

Beginning with (a), the clear sky (unstructured) is correctly delineated from the building (moving structure). The unstructured region represents a situation where there is insufficient information to determine flow. In (b), although the spatial appearance alone is insufficient to make the surface distinctions, the two surfaces are successfully partitioned based on their spacetime structure signatures. In (c), the stabilized foreground (static) is clearly delineated from the rightward moving background. In (d), although significant temporal aliasing is present due to the rapid motion of the leopard, the portion attributed to the leopard is successfully delineated from the static tree. An optical flow abstraction, although feasible, would require preprocessing the input with a stabilization procedure (e.g., pyramid processing [19]) to bring the speed of the leopard within its operating speed range ($\sim$ 1 pixel/frame). In (e), similar to (d), correct delineation is achieved in the presence of significant temporal aliasing. In (f), the sequence is successfully partitioned into two coherent regions, one consisting of a leftward motion while the second consisting of a scintillating region. Inspection of our oriented energy representation reveals a fairly uniform energy distribution at each point consistent with scintillation. Since the assumption of brightness conservation is clearly violated, optical flow would be inappropriate here. From a semantic viewpoint, the water group is consistent with one's impression that some coherent process is being imaged. In (g), the recovered dominant coherent structures consist of: unstructured (the wall), static (the painting) and flicker (the hallway). The painting is partially over-segmented along the boundaries due to the strong one-dimensional structure along the painting's boundaries that induce an oriented energy signature consistent with the aperture problem. At this level of analysis, this

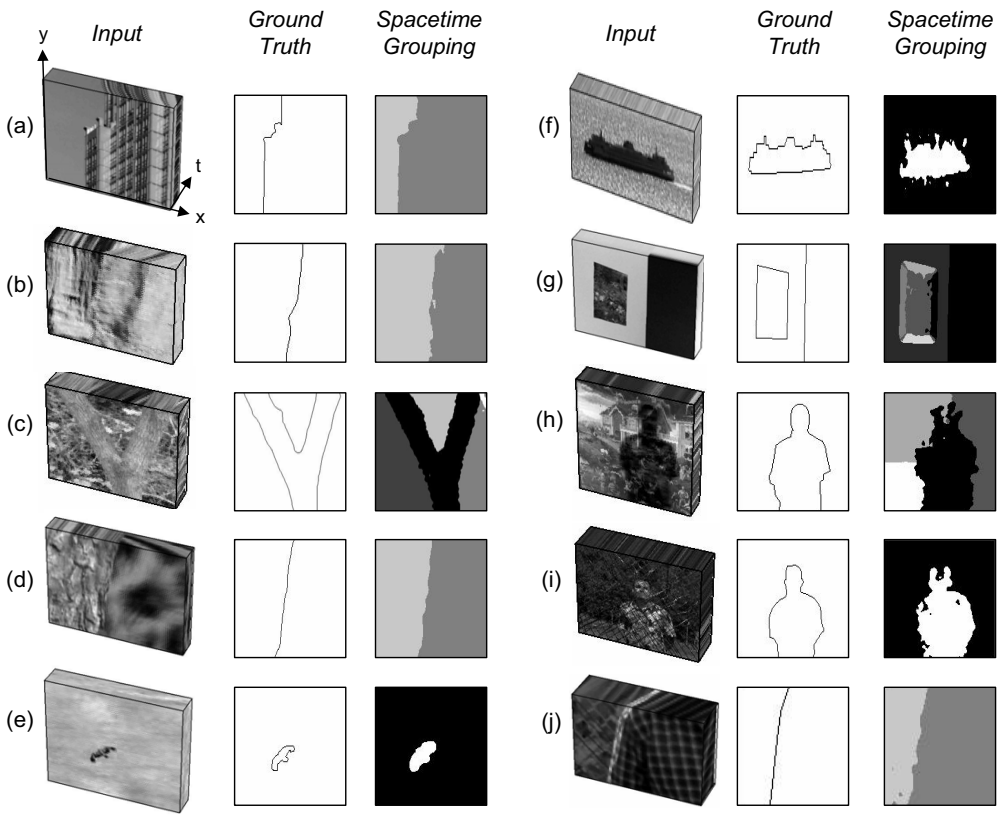

Figure 4. Successful grouping results on a diverse and challenging set of natural imagery. In each example, the input sequence, a frame from the human-labeled ground truth and grouping result, resp. are given. (a) A panning sequence consisting of a clear sky (i.e., unstructured) and a building (source: 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).

is a reasonable interpretation/grouping: The regions near the border of the painting against the unstructured wall are uniformly consistent as being characterized by the aperture problem. In (h), the translucent region (static background with superimposed moving person) is successfully distinguished from the purely static background, although the static background is slightly over-segmented due to low texture regions. Optical flow is inappropriate in the translucency region since it cannot be described by a single motion. In (i), the sequence is partitioned into the two constituent coherent structures: static structure vs. superimposed static and leftward motion (i.e., pseudo-transparency). Finally in (j), the sequence is partitioned into the two constituent coherent structures: rightward motion vs. superimposed static and rightward motion (i.e., pseudo-transparency). Notice, even with an oriented structure mutually shared in both patterns (rightward motion), successful segregation is achieved on the basis of the additional static orientation in one of the patterns due to the fence.

The grouping results in Fig. 4 provide compelling qualitative evidence that the proposed approach to grouping performs well on image sequences containing a variety of spatiotemporal structures. To quantify performance, results are compared with frame-by-frame human labeled (spatial) boundary ground truth. Following a standard evaluation methodology for image segmentation [13], mean precision/recall scores were calculated across all the image sequences illustrated in Fig. 4 and are shown as tuning curves in Fig. 5. These curves characterize the grouping performance over a range of the input parameters. *Precision* is

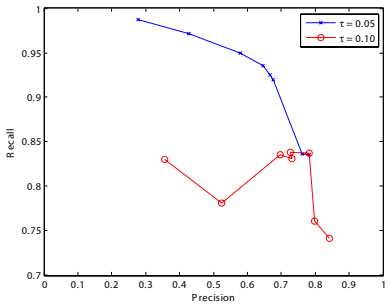

Figure 5. Precision/recall curves. Points along the curve correspond to variations of the range bandwidth within $[0.04, 0.18]$.

defined as the percentage of *detected* boundary pixels in the *recovered grouping*[2] that correspond to *ground truth* boundary pixels:

$$Precision = Matched(S_r, S_g)/|S_r|, \quad (7)$$

where S_r and S_g represent the grouping result and ground truth, resp., $Matched(S_r, S_g)$ denotes the number of boundary pixels in S_r that were matched in S_g within a neighbourhood of radius ρ and $|\cdot|$ denotes the cardinality of the set of boundary pixels. *Recall* is defined as the percentage of *ground truth* boundary pixels that were *detected* in the *recovered grouping*:

$$Recall = Matched(S_g, S_r)/|S_g|. \quad (8)$$

Over-segmentation is characterized in the curves by high recall but low precision, and the converse holds for under-segmented images.

For mean-shift, there are four input parameters, the bandwidths, $\{h_{space}, h_{time}, h_{range}\}$, and the merging threshold, τ. Through preliminary experimentation it was found that the largest differences between grouping results are realized when varying the range bandwidth and merging threshold parameters. In Fig. 5, points along the curve correspond to varying the range bandwidth, h_{range}, within $[0.04, 0.18]$ while fixing the spatial and temporal bandwidths to $(h_{space}, h_{time}) = (32, 10)$ and merging threshold to $\tau \in \{0.05, 0.10\}$. Matching was carried out using a distance threshold $\rho = 8$, which is reasonable given that the oriented energy filtering support also spans eight pixels. The consistently high recall indicates that the recovered groupings show little tendency to under-segment and thus one can conclude that the combination of representation and grouping methods captures salient image structure. At the same time, a relatively high precision is attained, which indicates that significant over-segmentation is not prevalent.

3.3. Additional examples

Figure 6 illustrates a successful grouping result on a synthetic input sequence consisting of a counter-clockwise rotating pink noise disc over a stationary pink noise background. This example indicates the approach's ability to recover groupings in cases where spacetime structure varies smoothly. A similar example has been successfully demonstrated using the tensor voting framework [21]; however, that approach relies on flow, while the examples of Fig. 3 show that flow yields difficulties in situations involving more complicated patterns (e.g., transparency and scintillation). Significantly, in the present framework the cases of grouping smoothly varying flow as well as non-motion dynamic structure can be handled in a uniform fashion.

[2]Boundaries in the labeled groupings are identified as any pixel that has at least a single neighbour with a differing label. This is followed by a thinning procedure to realize 1 pixel width boundaries.

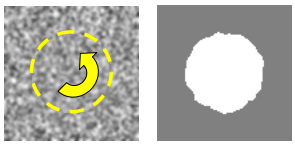

Figure 6. Grouping on a (synthetic) smoothly varying spacetime structure. (left) Input consisting of a counter-clockwise rotating pink noise disc over a stationary pink noise background. (right) A sample frame from the successful bipartite grouping result.

Figure 7. Grouping of a target initially at rest, then in motion. (left-to-right) Input frame of initially static scene; input frame taken after the onset of motion; grouping result of initial frame shows no foreground/background delineation as person and surround are both static; grouping result as person is in motion. Correct grouping is maintained throughout person's subsequent motion.

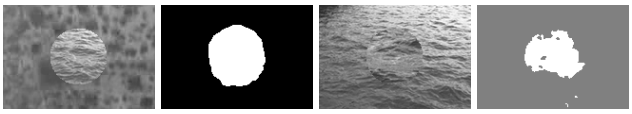

Figure 8. Grouping on a set of (synthetically combined) dynamic textures used in [4]. (left-to-right) Input frame of two widely differing dynamic textures (turbulent water and rising bubbles); grouping result using the proposed approach; input frame consisting of two instances of turbulent water moving left and right, respectively, both textures share the same appearance but differ in dynamics; grouping result using the proposed approach.

Figure 7 shows a successful grouping result on a natural input sequence, where spacetime structure changes over time, as a foreground target initially at rest goes into motion against a static background. This example demonstrates the approach's ability to deal with both spatial and temporally juxtaposed structure in a uniform manner.

Figure 8 shows successful grouping results on two dynamic texture examples [4]. Previously reported results on these examples require knowledge of the number of textures as well as hand contour initialization and show significant under-segmentation on the second example [4]. The present approach does not require similar a priori knowledge, yet yields comparable (first example) or superior (second example) results.

4. Discussion and summary

The main contribution of the presented research is the representation of visual spacetime via spatiotemporal orientation distributions so that diverse structures can be grouped readily into coherent spacetime regions. The work builds upon previous efforts on spacetime oriented filtering but differs markedly in that filtering is executed to capture the (possibly multi-) oriented structure of raw dynamic visual data rather than the recovery of an assumed flow field (i.e., the dominant spacetime orientation). Significantly, the same oriented representation supports further abstractions of visual spacetime data, both qualitative [26] and quantitative [22].

In contrast to the main theme of representation and grouping of spacetime information with a distributed representation of spatiotemporal orientation, the specific details of implementation are less important. In terms of oriented filtering, a variety of different filters could have been utilized, such as *higher-order directional cosine*, *Gabor*, *log-normal* and *causal-time* filters [19]. Similarly for grouping, mean-shift can be replaced by one of the alternatives cited in Sec. 1.2. Nevertheless, it is promising that the straightforward use of a popular grouping mechanism with a fairly simple set of single scale filters operating in a coarsely quantized orientation space have worked on a variety of natural image sequences.

In summary, this paper has presented a unified approach to representing and grouping a wide range of juxtaposed spacetime patterns (motion, static, flicker, (pseudo-) transparency, translucency, scintillation/temporal texture, unstructured). The approach is based on a distributed characterization of visual spacetime in terms of 3D, (x, y, t), spatiotemporal orientation. Empirical evaluation on a wide variety of imagery demonstrates the approach's ability to parse spacetime imagery into coherently structured regions. The delineated coherent regions, in conjunction with the underlying representation of spatiotemporal orientation, can serve as building blocks for further organization and interpretation of visual spatiotemporal information.

Acknowledgements

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

References

[1] S. Beauchemin and J. Barron. The computation of optical flow. *ACM Comp. Surv.*, 27, 1995.

[2] T. Brox, A. Bruhn, and J. Weickert. Variational motion segmentation with level sets. In *ECCV*, pages I: 471–483, 2006.

[3] K. Cannons and R. Wildes. Spatiotemporal oriented energy features for visual tracking. In *ACCV*, pages I: 532–543, 2007.

[4] A. Chan and N. Vasconcelos. Modeling, clustering, and segmenting video with mixtures of dynamic textures. *PAMI*, 30(5):909–926, May 2008.

[5] O. Chomat and J. Crowley. Probabilistic recognition of activity using local appearance. In *CVPR*, pages II: 104–109, 1999.

[6] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *PAMI*, 24:603–619, 2002.

[7] D. Cremers and S. Soatto. Variational space-time motion segmentation. In *ICCV*, pages 886–893, 2003.

[8] T. Darrell and E. Simoncelli. Separation of transparent motion into layers using velocity-tuned mechanisms. Technical report, MIT, VisMod TR-244, 1993.

[9] D. DeMenthon and R. Megret. Spatio-temporal segmentation of video by hierarchical mean shift analysis. Technical report, U. Maryland, 2002.

[10] K. Derpanis and J. Gryn. Three-dimensional nth derivative of Gaussian separable steerable filters. In *ICIP*, pages III: 553–556, 2005.

[11] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *PETS*, pages 65–72, 2005.

[12] G. Doretto, D. Cremers, P. Favaro, and S. Soatto. Dynamic texture segmentation. In *ICCV*, pages 1236–1242, 2003.

[13] F. Estrada and A. Jepson. Quantitative evaluation of a novel image segmentation algorithm. In *CVPR*, pages II: 1132–1139, 2005.

[14] C. Fowlkes, S. Belongie, and J. Malik. Efficient spatiotemporal grouping using the Nystrom method. In *CVPR*, pages I:231–238, 2001.

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

[16] G. Granlund and H. Knutsson. *Signal Processing for Computer Vision*. Kluwer, December 1995.

[17] H. Greenspan, J. Goldberger, and A. Mayer. Probabilistic space-time video modeling via piecewise GMM. *PAMI*, 26(3):384–396, 2004.

[18] D. Heeger. Model for the extraction of image flow. *JOSA-A*, 2(2):1455–1471, 1987.

[19] B. Jähne. *Digital Image Processing, sixth edition*. Springer-Verlag, Berlin, 2005.

[20] R. Nelson and R. Polana. Qualitative recognition of motion using temporal texture. *CVGIP*, 56(1):78–89, 1992.

[21] M. Nicolescu and G. Medioni. Layered 4D representation and voting for grouping from motion. *PAMI*, 25(4):492–501, 2003.

[22] E. Simoncelli. *Distributed Analysis and Representation of Visual Motion*. PhD thesis, MIT, 1993.

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

[24] S. Ullman. *The Interpretation of Visual Motion*. MIT Press, Cambridge, 1979.

[25] A. Watson and A. Ahumada. A look at motion in the frequency domain. In *Motion Workshop*, pages 1–10, 1983.

[26] R. Wildes and J. Bergen. Qualitative spatiotemporal analysis using an oriented energy representation. In *ECCV*, pages II: 768–784, 2000.