In *Proceedings of the European Conference on Computer Vision*, 768-784, 2000.

Qualitative Spatiotemporal Analysis Using an Oriented Energy Representation

Richard P. Wildes and James R. Bergen

Sarnoff Corporation
Princeton, NJ 08543
USA

Abstract. This paper presents an approach to representing and analyzing spatiotemporal information in support of making qualitative, yet semantically meaningful distinctions at the earliest stages of processing. A small set of primitive classes of spatiotemporal structure are proposed that correspond to categories of stationary, coherently moving, incoherently moving, flickering, scintillating and "too unstructured to support further inference". It is shown how these classes can be represented and distinguished in a uniform fashion in terms of oriented energy signatures. Further, empirical results are presented that illustrate the use of the approach in application to natural imagery. The importance of the described work is twofold: (i) From a theoretical point of view a semantically meaningful decomposition of spatiotemporal information is developed. (ii) From a practical point of view, the developed approach has the potential to impact real world image understanding and analysis applications. As examples: The approach could be used to support early focus of attention and cueing mechanisms that guide subsequent activities by an intelligent agent; the approach could provide the representational substrate for indexing video and other spatiotemporal data.

1 Introduction

1.1 Motivation

When confronted with spatiotemporal data, an intelligent system that must make sense of the ensuing stream can be overwhelmed by its sheer quantity. Video and other temporal sequences of images are notorious for the vast amount of raw data that they comprise. An initial organization which affords distinctions that can guide subsequent processing would be a key enabler for dealing efficiently with data of this nature.

The current paper explores the possibility of performing qualitative analyses of spatiotemporal patterns that capture salient and meaningful categories of structure and which are easily recovered from raw data. These categories capture distinctions along the following lines: What is moving and what is stationary? Are the moving objects moving in a coherent fashion? Which portions of the data are best described as scintillating and which portions are simply too unstructured to support subsequent analysis? More generally, given a spatiotemporal region of interest, one may seek to decompose it into a combination of such components. Significantly, it is shown that all of these distinctions can be based on a unified representation of spatiotemporal information in terms of local (spatiotemporal) correlation structure.

The ability to parse a stream of spatiotemporal data into primitive, yet semantically meaningful, categories at an early stage of analysis can benefit subsequent processing in a number of ways. A parsing of this type could support cueing and focus of attention for subsequent analysis. Limited computational resources could thereby be focused on portions of the input data that will support the desired analysis. For example, areas that are too unstructured to support detailed analysis could be quickly discarded. Similarly, appropriate models to impose during subsequent analysis (such as for model-based motion estimation) could be selected and initialized. Further, the underlying representation could provide the basis of descriptors to support the indexing of video or other spatiotemporal data. The relative distribution of a spatiotemporal region's total energy across the defined primitives might serve as a characteristic signature for initial database construction as well as subsequent look-up. Also, in certain circumstances the proposed analysis could serve directly to guide intelligent action relative to the impinging environment. Certain primitive reactive behaviors (say, pursuit or flight) might be triggered by the presence of certain patterns of spatiotemporal structure (say, patterns indicative of large moving regions). As a step toward such applications, this paper presents an approach to qualitative spatiotemporal analysis and illustrates its representational power relative to a variety of natural image sequences.

1.2 Related research

Previous efforts that have attempted to abstract qualitative descriptors of motion information are of relevance to the research described in the current paper. Much of this work is motivated by observations suggesting the inherent difficulty of dealing with the visual motion field in a quantitative fashion [27] as well as the general efficacy of using motion in a qualitative fashion to solve useful tasks (e.g., boundary and collision detection) [26]. It should be noted, however, that the focus of most of this work is the qualitative interpretation of *visual motion* or *optical flow* while the current paper is about the analysis of *spatiotemporal structure*. The level of processing discussed here precedes that at which actual motion computation is likely to occur. Indeed, one possible use of low-level spatiotemporal structure information might be to determine where optical flow computation makes sense to perform.

Recent advances in the use of parameterized models characterizing motion information in terms of its projection onto a set of basis flows are also of interest. Some of this work makes use of principle components analysis to build the basis flows from training data with estimation for new data based on searching the space of admissable parameters [5]. Other work has defined steerable basis flows for simple events (e.g., motion of occluding edge or bar) with subsequent ability to both detect and estimate weights for a novel data set [9]. As a whole, this body of research is similar to the previously reviewed qualitative motion analysis literature in being aimed at higher-level interpretation.

Most closely related to the current work is prior research that has approached motion information as a matter for temporal texture analysis [17]. This research is similar in its attempt to map spatiotemporal data to primitive, yet meaningful patterns. However, it differs in significant ways: Its analysis is based on statistics (e.g., means and variances) defined over normal flow recovered from image sequence intensity data; whereas, the current work operates directly on the intensity data. Further, the patterns that it abstracts to (e.g., flowing water, fluttering leaves) are more specific and narrowly defined than those of the current work.

A large body of research has been concerned with effecting the recovery of image motion (e.g., optical flow) on the basis of filters that are tuned for local spatiotemporal orientation [1, 8, 11–13, 28]. Filter implementations that have been employed to recover estimates of spatiotemporal orientation include angularly tuned Gabor, lognormal and derivative of Gaussian filters. Also of relevance is the notion of opponency between filters that are tuned for different directions of motion [1, 21, 23]. An essential motivation for taking such an operation into account is the close correspondence between the difference in the response of filters tuned to opposite directions of motion (e.g., leftward vs. rightward) and optical flow along the same dimension (e.g., horizontal). While the current work builds directly on methods for recovering local estimates of spatiotemporal orientation, it then takes a different direction in moving directly to qualitative characterization of structure rather than the computation of optical flow.

Previous work also has been concerned with various ways of characterizing local estimates of spatiotemporal orientation. One prominent set of results along these lines has to do with an eigenvalue analysis of the local orientation tensor [11, 14]. Here the essential point is to characterize the dimensionality of the local orientation as being isotropic, line- or plane-like in order to characterize the local spatial structure with respect to motion analysis (e.g., distributed vs. oriented spatial structure with uniform motion). Other work of interest along these lines includes interpretation of opponent motion operators as indicative of motion salience [30] and the exploitation of multiscale analysis of temporal change information for detection and tracking purposes [2]. Overall, while these lines of investigation are similar to the subject of the current paper, none of this work has proposed and demonstrated the particular and complete set of spatiotemporal abstractions that are the main subject of the current paper.

In the light of previous research, the main contribution of the current paper is that it shows how to abstract from spatiotemporal data a number of qualitative structural descriptions corresponding to semantically meaningful distinctions (e.g., what is stationary, what is moving, is the exhibited motion coherent or not, etc.). Further, a formulation is set forth that captures all of the distinguished properties of spatiotemporal structure in a unified fashion.

2 Technical approach

In this section, the proposed approach to spatiotemporal analysis is presented, accompanied by natural image examples. For the purposes of exposition, the presentation begins by restricting consideration to one spatial dimension plus time. Subsequently, the analysis is generalized to encompass an additional spatial dimension and issues involving spatiotemporal boundaries.

2.1 Analysis in one spatial dimension plus time

	Unstructured	Static	Flicker	Coherent Motion	Incoherent Motion	Scintillation
$\|R-L\|$	0	0	0	++	0	0
$R+L$	0	++	++	++	+ + ++	++
S_x	0	++	0	+	+	+
F_x	0	0	++	+	+	+

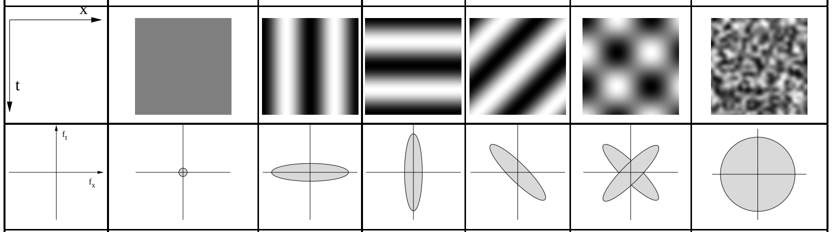

Fig. 1. Primitive Spatiotemporal Patterns. The top row of images depict prototypical patterns that comprise the proposed qualitative categorization of spatiotemporal structure. For display purposes the images are shown for a single spatial dimension, x, plus time, t. The second row of plots shows the corresponding frequency domain structure, with axes f_x and f_t. As suggested by their individual titles, the categories have semantically meaningful interpretations. The lower part of the figure shows the predicted distribution of energy for each pattern as it is brought under the proposed oriented energy representation. The representation consists of four energy images components, $|R-L|$, $|R+L|$, S_x and F_x that are derived from an input image via application of a bank of oriented filters. For the purpose of qualitative analysis the amount of energy that is contributed by the underlying filter responses, R, L, S_x and F_x, is taken as having one of three values: (approximately) zero, moderate and large, symbolized as 0, + and ++, respectively.

Primitive spatiotemporal patterns The local orientation (or lack thereof) of a pattern is one of its most salient characteristics. From a purely geometric point of view, orientation captures the local first-order correlation structure of a pattern. In the realm of image analysis, local spatiotemporal orientation often can be interpreted in a fashion that has additional ramifications. For example, image velocity is manifest as orientation in space-time [14]. We now explore the significance of this structure in one spatial dimension, the horizontal image axis,

x, and time, t. Fig. 1 shows x-t-slices of several prototypical spatiotemporal patterns that are of particular interest.

Perhaps the simplest situation that might hold is that a region is essentially devoid of structure, i.e., image intensity is approximately constant or slowly varying in both the spatial and temporal directions. In the spatiotemporal frequency domain, such a pattern would have the majority of its energy concentrated at the origin. When such regions occur where local contrast is small they can indicate an underlying smoothness in the material that is being imaged. For subsequent processing operations it is important to flag such areas as lacking enough information to support stable estimates of certain image properties. For example, image registration can be led astray by blindly attempting to align structureless regions. This category will be referred to as "unstructured".

Locally oriented structures are quite common in spatiotemporal data. Here, there are several situations that are useful to distinguish. From a semantic point of view, it is of particular interest to categorize the patterns according to the direction of their dominant orientation. One case of interest is that which arises for the case of (textured) stationary objects. These cases show elongated structure in the spatiotemporal domain that is parallel to the temporal axis, i.e., features exhibit no shift in position with the passage of time. In the frequency domain, their energy will be concentrated along the spatial frequency axis. This case will be referred to as "static". A second case of interest is that of homogeneous spatial structure, but with change in intensity over time (for example, overall change in brightness due to temporal variation in illumination). Here, the spatiotemporal pattern will be oriented parallel to the spatial axis. Correspondingly, in the frequency domain the energy will be concentrated along the temporal frequency axis. This case will be referred to as "flicker". A third case of interest is that of objects that are in motion. As noted above, such objects trace a trajectory that is slanted in the spatiotemporal domain in proportion to their velocity. Their energy in the frequency domain also exhibits a slant corresponding to their having both spatial and temporal variation. Such simple motion that is (at least locally) manifest by a single dominant orientation will be referred to as "coherent motion". Finally, it is useful to distinguish a special case of oriented structure, that of multiple local orientations intermixed or superimposed within a spatial region. In this regard, there is motivation to concentrate on the case of two structures both indicative of motion. Such a configuration has perceptual significance corresponding to oscillatory motion, shear and occlusion boundaries, and other complex motion phenomena that might be generally thought of as dynamic local contrast variation with motion. Interestingly, it appears that human vision represents this category as a special case as suggested by the perception of counterphase flicker [6]. In the frequency domain the energy distribution will be the sum of the distributions that are implied by the component motions. This case will be referred to as "incoherent motion". In comparison, there does not seem to be anything significant about something that is both static and flickering, beyond its decomposition into those primitives.

The final broad class of spatiotemporal pattern to be considered is that of isotropic structure. In this case, no discernable orientations dominate the local region; nevertheless, there is significant spatiotemporal contrast. The frequency domain manifestation of the pattern also lacks a characteristic orientation, and is likewise isotropic. Situations that can give rise to this type of structure are characteristically stochastic or chaotic in nature. Natural examples include turbulence and the glint of specularities on water. Owing to the perceptual manifestation of these phenomena, this case will be referred to as "scintillation".

The essence of the proposed approach is to analyze any given sample of spatiotemporal data as being decomposed along the dimensions of the adduced categories: unstructured, static, flicker, coherent motion, incoherent motion and scintillation. While it is possible to make finer distinctions (e.g., exactly what the numerical value of the space-time orientation is), at the level of qualitative semantics these are fundamental distinctions to be made: Is something structured or not? If it is structured, does it exhibit a characteristic orientation or is it more isotropic and thereby scintillating in nature? Are oriented patterns indicative of something that is stationary, flickering or moving? Is the motion coherent or incoherent? It should be noted that each of the descriptions identified above is attached to the visual signal within a specified spatiotemporal region. The choice of this region generally affects the description assigned. For example, the motion of leaves in the wind may be coherent if analyzed over a very small area and time but incoherent over a larger area or time. An alternative way to think about the proposed decomposition is to consider it from the point of view of signal processing: In particular, what sort of decomposition (e.g., in the frequency domain) does it imply. This topic is dealt with in the next section in terms of a representation that captures the proposed distinctions.

Oriented energy representation Given that the concern is to analyze spatiotemporal data according to its local orientation structure, a representation that is based on oriented energy is appropriate. Such a representation entails a filter set that divides the spatiotemporal signal into a set of oriented energy bands. In general, the size and shape of the filter spectra will determine the way that the spatiotemporal frequency domain is covered. In the present case, a family of relatively broadly tuned filters is appropriate due the interest in qualitative analysis. The idea is to choose a spatial frequency band of interest with attendant low pass filtering in the temporal domain. This captures orientation orthogonal to the spatial axis. On the basis of this choice, a temporal frequency band can be specified based on the range of dynamic phenomena that are of interest for the given spatial band. This captures structure that is oriented in directions indicative of motion, e.g., a spatiotemporal diagonal. Finally, these characteristics can be complemented by considering just the temporal frequency band while spatial frequency is covered with a low-pass response. This captures structure that is oriented orthogonal to the temporal axis. Thus, it is possible to represent several principle directions in the spatiotemporal domain while systematically covering the frequency domain. The simplification realized by analyzing

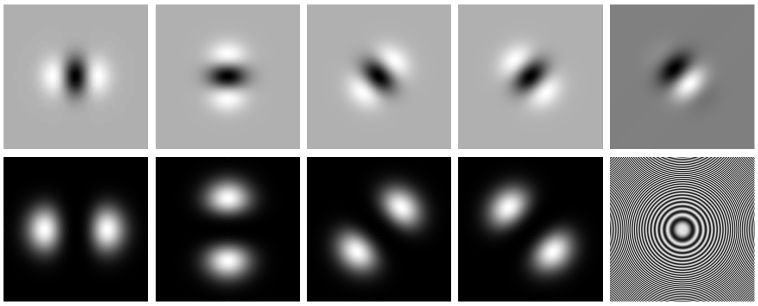

Fig. 2. Oriented Energy Filters for Spatiotemporal Analysis. The top row shows synthesized profiles for second derivative of Gaussian filters oriented to capture static, flicker, rightward and leftward motion structure (left to right). The last plot is the Hilbert transform of the leftward motion filter. (These plots are shown greatly enlarged for clarity). The bottom row indicates the frequency response of the corresponding quadrature pair filters via application of an energy calculation to the zone plate at the far right. The proposed approach to representing spatiotemporal structure builds on such filtering operations.

spatiotemporal structure in a two dimensional representation (i.e. one spatial and one temporal dimension) requires somehow addressing the remaining spatial dimension since the input data consists of a three dimensional volume. This is done by lowpass filtering the data in the orthogonal spatial direction using the 5-tap binomial filter $[1\,4\,6\,4\,1]/16$. This filtering allows for analysis of the other spatiotemporal plane (i.e. that containing the orthogonal spatial dimension) in an exactly analogous manner.

In the remainder of this section a choice of filters is presented for a given frequency response, i.e., scale of spatial structure.

The desired filtering can be implemented in terms of second derivative of Gaussian filters, G_{2_θ} at orientation θ (and their Hilbert transforms, H_{2_θ}) [14]. The motivation for this choice is twofold. First, while selective for orientation, the tuning of these filters is moderately broad and therefore well suited to the sort of qualitative analysis that is the focus of the current research. Second, they admit a steerable and separable implementation that leads to compact and efficient computation. The filters are taken in quadrature (i.e., for any given θ, G_{2_θ} and H_{2_θ} in tandem) to eliminate phase variation by producing a measure of local energy, $E_\theta(x,t)$ within a frequency band, according to

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

In particular, to capture the principle orientations that were suggested above, filtering is applied (i) oriented orthogonally to the spatial axis, (ii) orthogonally to the temporal axis and (iii, iv) along the two spatiotemporal diagonals, see Fig. 2. Notice that the frequency response plots show how the filters sweep out an annulus in that domain; this observation can provide the basis for allowing a multiscale extension to systematically alter the inner and outer rings of the annulus to effectively cover the frequency domain. Finally, note that at a given frequency the value of any one oriented energy measure is a function of both orientation and contrast and therefore rather ambiguous. To avoid this confound and get a purer measure of orientation the response of each filter should be normalized by the sum of the consort, i.e.,

$$\hat{E}_{\theta_i}(x,t) = \frac{E_{\theta_i}}{\Sigma_i E_{\theta_i(x,t)} + \epsilon} \tag{2}$$

where ϵ is a small bias to prevent instabilities when overall energy is small. (Empirically we set this bias to about 1 % of the maximum (expected) energy.)

The necessary operations have been implemented in terms of a steerable filter architecture [10, 15]. The essential idea here is to convolve an image of interest with a set of n basis filters, with $n = 3$ for the second derivative of Gaussians of concern. Subsequently the basis filtered images are combined according to interpolation formulas to yield images filtered at any desired orientation, θ. Processing with the corresponding Hilbert transforms is accomplished in an analgous fashion, with $n = 4$. To remove high frequency components that are introduced by the squaring operation in forming the energy measurement (1), the previously introduced 5-tap binomial low-pass filter is applied to the result, E_θ. Details of the filter implementation (e.g., specification of the basis filters and the interpolation formulas) are provided in [10, 15].

The final oriented energy representation that is proposed is based directly on the basic filtering operations that have been described. Indeed, given the class of primitive spatiotemporal patterns that are to be distinguished, one might imagine simply making use of the relative distribution of (normalized) energies across the four proposed orientation tuned bands as the desired representation. In this regard, it is proposed to make use of two of these bands directly. In particular, the result of filtering an input image with the filter oriented orthogonally to the spatial axis will be one component of the representation, let it be called the "S_x-image" (for static). Second, let the result of filtering an input image with the filter oriented orthogonally to the temporal axis be the second component of the representation and call it the "F_x-image" (for flicker). Due to their characteristic highlighting of particular orientations, these (filtered) images are well suited to capturing the essential nature of the patterns for which they are named.

The information provided individually by the remaining two bands is ambiguous with respect to the desired distinctions between, e.g., coherent and incoherent motion. This state of affairs can be remedied by representing these bands as summed and differenced (i.e., opponent) combinations. Thus, let $R-L$ and $R+L$ stand for opponent and summed images (resp.) formed by taking the pointwise arithmetic difference and sum of the images that result from filtering an input image with the energy filters oriented along the two diagonals. It can be shown that the opponent image (when appropriately weighted for contrast) is proportional to image velocity [1] and has a strong signal in areas of coherent motion. It is for this reason that the notation R and L is chosen to underline the relationship to rightward and leftward motion. For present purposes the absolute value of the opponent signal, $|R-L|$, will be taken as the third component of the proposed representation since this allows for coherency always to be positive. Finally, the fourth component of the representation is the summed (motion) energy $R + L$. This image is of importance as it captures energy distributions that contain multiple orientations that are individually indicative of motion and is therefore of importance in dealing with incoherent motion phenomena.

At this point it is interesting to revisit the primitive spatiotemporal patterns of interest and see how they project onto the four component oriented energy representation comprised of S_x, F_x, $|R - L|$ and $R + L$, see Fig. 1. In the unstructured case, it is expected that all of the derived images will contain vanishingly small amounts of energy. Notice that for this to be true and stable, the presence of the bias factor, ϵ, in the normalization process is important in avoiding division by a very small factor. For the static case, not surprisingly the S_x-image contains the greatest amount of energy. Although, there also is a moderate energy from the $R + L$-image as the underlying R and L responses will be present due to the operative orientation tuning. In contrast, these responses will very nearly cancel to leave the $|R - L|$-image essentially zero. Similarly, the orthogonal F_x-image should have essentially no intensity. The flicker case is similar to the static case, with the S_x and F_x-images changing roles. For the case of coherent motion, it is expected that the $|R - L|$-image will have a large amount of energy present. Indeed, this is the only pattern where the image is expected to contain any significant energy. The $R + L$-image also should show an appreciable response, with the other images showing more moderate responses. For the case of incoherent motion, the $R + L$- image should dominate as both the underlying R and L responses should be appreciable. Again, due to finite bandwidth tuning the S and F images also should show moderate responses. Once again the $|R - L|$-image should be very nearly zero. Finally, for the case of scintillation the S and F images should show modest, yet still appreciable responses. The $R + L$-image response should be somewhat larger, perhaps by a factor of two as each of the modest R and L responses sum together. Essentially no response is expected from the $|R - L|$-image. Significantly, when one compares all of the signatures, each is expected to be distinct from the others, at least for the idealized prototypical patterns. The question now becomes how well the representation captures the phenomena of interest in the face of natural imagery.

Natural image examples A set of natural image sequences have been gathered that provide one real world example of each of the proposed classes of spatiotemporal structure, see Fig. 3. For the unstructured case the image sequence shows a featureless sky. For the static case the image sequence shows a motionless tree. (Note that for each of these first two cases, a single image was not simply duplicated multiple times to make the sequence, an actual video sequence of images was captured.) The third case, flicker, is given as a smooth surface of human skin as lightning flashes over time. Coherent motion is captured by a field of flowers that appear to move diagonally upward and to the right due to camera motion. Incoherent motion is captured by a sequence of images of overlapping legs in very complex motion (predominantly, but not entirely, horizontal motion). The last case, scintillation, is shown via a sequence of rain striking a puddle. All of the image sequences had horizontal, x, and vertical, y, length both equal to 64 while the temporal length (i.e., number of frames) was 40. All of the spatiotemporal image volumes were processed in an identical fashion by bringing them under the proposed oriented energy representation, as described in the previous section. This resulted in each original image begin decomposed along the four dimensions, $|R-L|$, $R+L$, S_x and F_x.

The results of the analysis are shown in Fig. 3. For each of the natural image examples a representative spatial slice shows the recovered energy along each of the dimensions, $|R-L|$, $R+L$, S_x and F_x. In each cell, the average (normalized) energy is shown for the entire spatiotemporal volume. (Note that due to the presence of the bias, ϵ, the sum of $R+L$, S_x and F_x does not necessarily sum exactly to unity.) In reviewing the results it is useful to compare the recovered distribution of energies with the predictions that are shown in Fig. 1. Beginning with the unstructured case, it is seen that all of the recovered energies are vanishingly small, exactly as predicted. The static case also follows the pattern predicted in Fig. 1. For this case it is interesting to note that the deviation from zero in the F_x component is due to some fluttering (i.e., scintillation) in the leaves of the tree. The flicker case also performs much as expected, with a bit more energy in the F_x component relative to the $R+L$ component than anticipated. For the case of coherent motion the pattern of energy once again follows the prediction closely. Here it is important to note that the depicted motion is not strictly along the horizontal axis, rather it is diagonal. This accounts for the value of $R+L$ being somewhat larger than $|R-L|$ as the underlying L channel has a nonzero response. For the incoherent case, it is seen that while the general trend in the distribution of energies is consistent with predictions, the magnitude of $R+L$ is not as large as expected. Examination of the data suggests that this is due to the F_x component taking on a larger relative value than expected due to the imposition of some flicker in the data as some bright objects come into and go out of view (e.g., bright props and boots that the people wear). Finally, the case of scintillation follows the predictions shown in Fig. 1 quite well. Taken on the whole, these initial empirical results support the ability of the proposed approach to make the kinds of distinctions that have been put forth. Clearly the utility of the representation depends on its ability to distinguish and identify populations of samples corresponding to the various semantic categories described. Demonstration of this ability will require a quantitative analysis of energy signatures across an appropriate collection of samples and is beyond the scope of this paper.

2.2 Adding an additional spatial dimension

The approach that has been developed so far can be extended to include the vertical dimension, y, by augmenting the representation with a set of components that capture oriented structure in y-t image planes. The same set of oriented filters that were used previously are now applied to y-t planes, as before with the addition of a low-pass characteristic in the orthogonal spatial dimension, now x. This will allow for (normalized) oriented energy to be computed in the four directions: (i) oriented orthogonally to the spatial axis, y, (ii) oriented orthogonally to the temporal axis, t and (iii,iv) along the two y-t diagonals. These energy computations are performed for an input image using the y-t counterparts of formulas (1) and (2). The resulting filtered images are then used to complete the representation in a way entirely analogous to that used for the horizontal case except with U and D (for up and down) replacing R and L.

To illustrate these extensions, Fig. 4 shows the results of bringing the same set of natural image examples that were used with the x-t analysis under the $|U-D|$, $U+D$, S_y, F_y extensions to the representation. Here it is useful to refer to both the a priori predictions of Fig. 1 as well as the previously presented x-t empirical results. By and large the results once again support the ability of the approach to distinguish the six qualitative classes that have been put forth. Note, however, that for the incoherent motion case the depicted movement is predominant in the x direction and the value of $U+D$ is correspondingly relatively low.

2.3 Boundary analysis

As an example of how the proposed representation can be used for early segmentation of the input stream, we consider the detection of spatiotemporal boundaries. Differential operators matched to the juxtaposition of different kinds of spatiotemporal structure can be assembled from the primitive filter responses, $R-L$, $R+L$, S_x, F_x and their vertical (i.e., y-t) counterparts. To illustrate this concept, consider the detection of (coherent) motion boundaries. Here, the intent is not to present a detailed discussion of motion boundary detection, which has been extensively treated elsewhere (see, for example [3, 7, 9, 19]), but to use it as an example of the analysis of spatiotemporal differential structure in general.

Coherent motion is most directly related to the opponent filtered images $R-L$ and $U-D$. Correspondingly, the detection of coherent motion boundaries is based on the information in these images. As shown in Fig. 5, combining a difference of Gaussians

$$G(x,y,\sigma_1) - G(x,y,\sigma_2) \tag{3}$$

operator (where $G(x,y,\sigma)$ is a Gaussian distribution with standard deviation σ) with motion opponent signals yields a double opponency: The pointwise opponency $R-L$ is combined with a spatial opponency provided by the difference of Gaussians and similarly for $U-D$. As in difference of Gaussian based edge-detection [14], the zero-crossings in the convolution of (3) with $R-L$ and $U-D$ are indicative of boundaries in these inputs. Final boundary detection is based on the presence of a zero-crossing in either of the individual results $(G(x,y,\sigma_1) - G(x,y,\sigma_2)) * (R-L)$ or $(G(x,y,\sigma_1) - G(x,y,\sigma_2)) * (U-D)$.

An example is shown in Fig. 5. Here, the difference of Gaussians (3) was realized in terms of binomial approximations to low-pass filters with cut-off frequencies at $\pi/8$ and $\pi/16$. A sequence of aerial imagery showing a tree canopy with movement relative to undergrowth due to camera motion serves as input.

Due to the homogeneous texture of the vegetation, the boundary of the tree is not visible in any one image from the sequence. Opponent motion images $R-L$ and $U-D$ were derived from this input and difference of Gaussian processing was applied to each of the motion opponent images. Finally, the zero-crossings in the results are marked. For purposes of display, the slope magnitude is calculated for the zero-crossings and summed between the two (double opponent) images to given an indication of the strength of the boundary signal. The result accurately captures one's visual impression upon viewing the corresponding image sequence where the apparent boundary can be traced along the left side as an irregular contour, then along a diagonal and finally across the top where it has a pronounced divot.

3 Discussion

3.1 Implications

The work that has been described in this paper builds on a considerable body of literature on spatiotemporal filtering. The main implication of the current effort is that the output of such filtering can be interpreted directly in terms of rather abstract information, i.e., the 6 proposed categories of spatiotemporal structure: structureless, static, flicker, coherent motion, incoherent motion and scintillation. Based on the analysis presented, not all of these classes are equally discriminable under the proposed representation. The signatures for the classes structureless, static, flicker and coherent motion are quite distinct, but those for incoherent motion and scintillation (while distinct from the other four) differ from each other only in the amount of energy expected in the summed energies $R+L$ and $U+D$. This state of affairs suggests that these last two categories might be best distinguished from each other in relative comparisons, while all other distinctions might be accomplished in a more independent and absolute fashion. This difference has implications for how the signatures can be used: The stronger form of distinctness supports categorical decisions about signal content across imaging situations; because it depends on a metric comparison, the weaker form probably does not.

Operations have been described at a single spatiotemporal scale; however, the proposed representation is a natural candidate for multiscale extensions [16, 31]. Indeed, such extensions might support finer distinctions between categories of spatiotemporal structure as characteristic signatures could be manifest across scale. Two kinds of extension can be distinguished. The first is concerned with varying the region of (spatiotemporal) integration that is applied to the oriented energy measures. The second type of multiscale extension concerns the frequency tuning of the underlying oriented filters. A systematic extension in this regard would operate at a number of spatial frequency bands and, for each of these bands, perform the analysis for a number of temporal frequency bands. It would thereby be possible to tile the frequency domain and correspondingly characterize the local orientation structure of an input spatiotemporal volume. These two extensions serve distinct purposes that are perhaps best understood with respect to a simple example. Consider a typically complex outdoor scene containing a tree blowing in a gusty wind and illuminated by a sunny sky with a few drifting clouds in it. As the tree branches sway back and forth, the corresponding image motion will be locally and temporarily coherent. However, over longer periods of time or over larger areas it will be incoherent or oscillatory. Thus, the characterization of the spatiotemporal structure will shift from one category to the other as the region of analysis is extended. Now consider the effect of a cloud shadow passing across the tree. At a fine spatial scale (i.e. for a high spatial frequency underlying filter) it will look like an illumination variation thus having a component in the "flicker" category. At the scale of the shadow itself (i.e. at low spatial frequency) it will look like coherent motion as the cloud passes over. The pattern of spatiotemporal signatures taken as a function of scale thus captures both the structural complexities of the dynamic scene and the quasitransparency of complex illumination. These two types of scaling behavior are complimentary and taken in tandem serve to enrich the descriptive vocabulary of the approach.

In contrast to the main message of this paper regarding the abstraction of spatiotemporal information to the level of qualitative descriptors, the details of the particular filtering architecture that have been employed are less important. A variety of alternatives could be employed, including oriented Gabor (e.g., [13]) and lognormal (e.g., [11]) filters. Similarly, one might be concerned with issues of causality and use oriented spatiotemporal filters that respect time's arrow [1, 8, 28]. Also, one might consider a more uniform sampling of orientation in x-y-t-space, rather than relying on x-t and y-t planes. Nevertheless, it is interesting that the fairly simple filters that were employed in the current effort have worked reasonably well for a variety of natural image examples.

The type of qualitative analysis described here seems particularly suited to processing in biological vision systems because of the apparently hierarchical nature of biological computation and the importance of such factors as attention. It is interesting therefore to note aspects of biological processing that relate to the current approach. With respect to fineness of sampling in the spatiotemporal domain, it appears that humans employ only about 2 to 3 temporal bands, while making use of 6 or more spatial bands [4, 25, 29]. Also, there is evidence that biological systems combine motion tuned channels in an opponent fashion [24], as is done in the current work. Regarding the degree to which filter responses are spatially integrated (i.e., as part of computing aggregate properties of a region) biological systems seem to be rather conservative: Physiological recordings of visual cortex complex cells indicate integration regions on the order of 2 to 5 cycles of the peak frequency [20], suggesting a preference for preservation of spatial detail over large area summation. It also is interesting to note that human contrast sensitivity is on the order of 1 % [18], an amount that has proven useful analogously in the current work as a choice for the bias in the process of energy normalization (2). With regard to border analysis, part of a purported mechanism for the detection of relative movement in the fly makes use of spatially antagonistic motion comparisons [22], in a fashion suggestive of the approach taken in the current paper.

Based on the ideas of this paper, a number of applications can be envisioned falling into two broad areas of potential impact. The first type of application concerns front end processing for real-time vision tasks. In this capacity, it could provide an initial organization, thereby focusing subsequent processing on portions of the data most relevant to critical concerns (e.g., distinguishing static, dynamic and low information regions of the scene). The second type of application concerns issues in the organization and access of video sequences. Here, the proposed representation could be used to define feature vectors that capture volumetric properties of spatiotemporal information (e.g., space-time texture) as an aid to the design and indexing of video databases. More generally, the proposed approach would be appropriate to a variety of tasks that could benefit from the early organization of spatiotemporal image data.

3.2 Summary

This paper has presented an approach to representing and analyzing spatiotemporal data in support of making qualitative yet semantically meaningful distinctions. In this regard, it has been suggested how to ask and answer a number in simple, yet significant questions, such as: Which spatiotemporal regions are stationary? Which regions are moving in a coherent (or incoherent) fashion? How much of the variance in the spatiotemporal data is due to overall changes in intensity. Where is the spatiotemporal structure isotropic and indicative of scintillation? Where is the data stream simply lacking in sufficient structure to support further inference? Also indicated has been an approach to issues regarding the analysis of spatiotemporal boundaries. Further, all of these matters have been embodied in a unified oriented energy representation. A variety of empirical results using natural image data suggest that the approach may have the representational power to support the desired distinctions. On the basis of these results, it is conjectured that the developed representation and analysis can subserve a variety of vision-based tasks and applications. More generally, the approach provides an integrated framework for dealing with spatiotemporal data in terms of its abstract information content at the earliest stages of processing.

References

1. Adelson, E., Bergen, J.: Spatiotemporal energy models for the perception of motion. JOSA A **2** (1985) 284–299
2. Anderson, C., Burt, P., van der Wal, G.: Change detection and tracking using pyramid transform techniques. Proc. SPIE Conf. on Intell. Rob. and Comp. Vis. (1985) 300–305
3. Beauchemin, S., Barron, J.: The computation of optical flow. ACM Comp. Surv. **27** (1995) 433-467
4. Bergen, J., Wilson, H.: Prediction of flicker sensitivities from temporal three pulse data. Vis. Res. (1985) 284–299

5. Black, M., Yacoob, Y., Jepson, A., Fleet, D.: Learning parameterized models of image motion. Proc. IEEE CVPR (1997) 561–567
6. Bruce, V., Green, B., Georgeson, M.: Visual Perception. East Sussex: Earlbaum (1996)
7. Chou, G.: A model of figure-ground segregation from kinetic occlusion. Proc. ICCV (1995) 1050–1057
8. Fleet, D., Jepson, A.: A Cascaded Approach to the Construction of Velocity Selective Mechanisms. RBCV Tech. Rep., TR-85-6, Dept. of Comp. Sci., University of Toronto (1985)
9. Fleet, D., Black, M., Jepson, A.: Motion feature extraction using steerable flow fields. Proc. IEEE CVPR (1998) 274–281
10. Freeman, A., Adelson, E.: The design and use of steerable filters. IEEE PAMI **13** (1991) 891–906
11. Granlund, G., Knutsson, H.: Signal Processing for Computer Vision. Boston: Kluwer (1995)
12. Grzywacz, N., Yuille, A.: A model for the estimation of local velocity by cells in the visual cortex. Proc. Roy. Soc. Lond. B **239** (1990) 129–161
13. Heeger, D.: A model for the extraction of image flow. JOSA A **4**, (1997) 1455–1471
14. Jähne, B.: Digital Image Processing. Berlin: Springer-Verlag (1993)
15. Knutsson, H., Wilson, R., Granlund, G.: Anisotropic non-stationary image estimation and its applications – part I: Restoration of noisy images. IEEE TC **31** (1983) 388–397
16. Koenderink, J.: Scale-time. Bio. Cyb. **58** (1988) 159–162
17. Nelson, R., Polana, R.: Qualitative recognition of motion using temporal texture. CVGIP-IU **56** (1992) 78–89
18. van Ness, R., Bouman, M.: Spatial modulation transfer in the human eye. JOSA **57** (1967) 401–406
19. Niyogi, S.: Detecting kinetic occlusion. Proc. ICCV (1995) 1044–1049
20. Movshon, J., Thompson, I., Tolhurst, D.: Receptive field organization of complex cells in the cat's striate cortex. J. Physiol. Lond. **283** (1978) 79–99
21. Reichardt, W.: Autocorrelation, a principle for the evaluation of sensory information by the central nervous system. In W. Rosenblith (Ed.) Sensory Communication, NY: Wiley (1961)
22. Reichardt, W., Poggio, T.: Figure-ground discrimination by relative movement in the visual system of the fly. Bio. Cyb. **35** (1979) 81-100
23. van Santen, J., Sperling, G.: Temporal covariance model of human motion perception. JOSA A **1** (1984) 451–473
24. Stromeyer, C., Kronauer, R., Madsen, J., Klein, S.: Opponent mechanisms in human vision. JOSA A **1** (1984) 876–884
25. Thompson, P.: The coding of the velocity of movement in the human visual system. Vis. Res. **24** (1984) 41–45
26. Thompson, W., Kearney, J.: Inexact vision. Proc. Workshop on Motion Rep. and Anal. (1986) 15–22
27. Verri, A., Poggio, T.: Against quantitative optical flow. IEEE PAMI **9** (1987) 171–180
28. Watson, A., Ahumada, A.: Model of human motion sensing. JOSA A **2** (1985) 322–341
29. Watson, A., Robson, J.: Discrimination at threshold: Labelled detectors in human vision. Vis. Res. **21** (1981) 1115–1122
30. Wildes, R.: A measure of motion salience. Proc. IEEE ICIP (1988) 183–187
31. Witkin, A.: Scale-space filtering. Proc. IJCAI (1983) 1019–1021

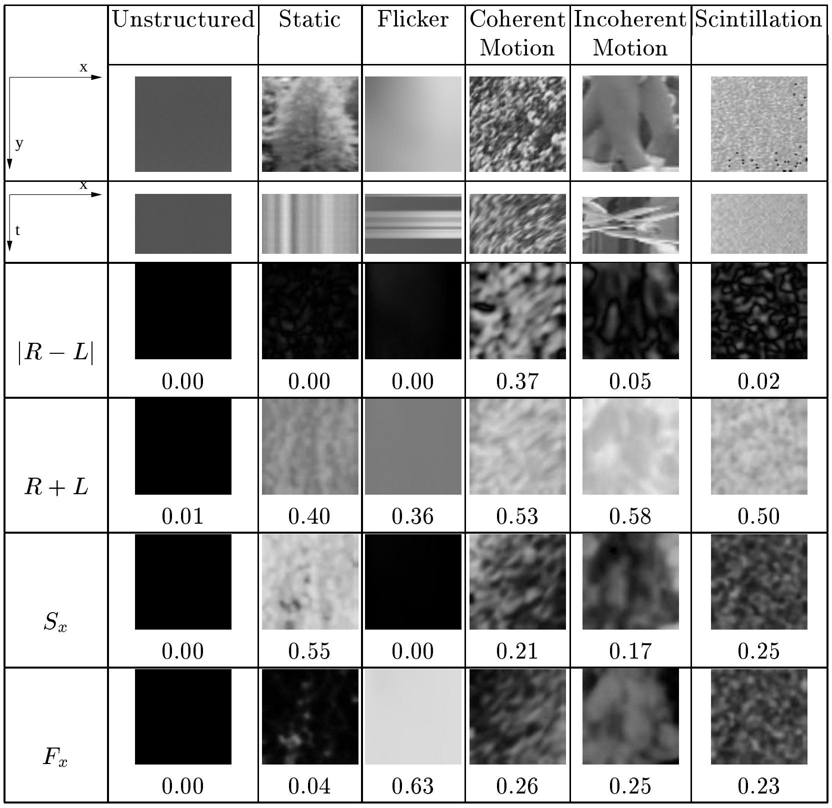

Fig. 3. Results of Testing the Proposed Representation on Natural Imagery. For each of the proposed primitive classes, a sequence of images that displays the associated phenomena was selected. Top row, left to right: featureless sky, a motionless tree, lightning flashing on (motionless) skin, a field of flowers in diagonal motion due to camera movement, legs of multiple cheerleaders in overlapping motion and rain striking a puddle. Each sequence has x, y, t dimensions of 64, 64, 40, respectively. The second row shows corresponding x-t-slices. The next four rows show the recovered energies in each of four components of the representation. Each cell shows a representative spatial, i.e., x, y, slice of the processed data as well as the average value for the energy across the entire spatiotemporal volume. Overall, the results are in accord with the predictions of Fig. 1.

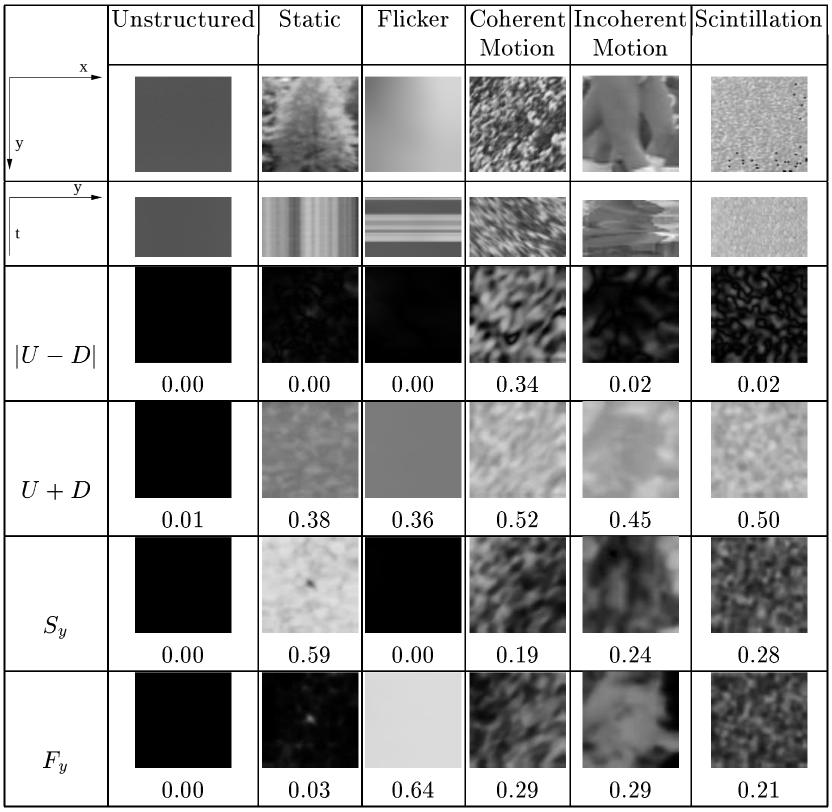

Fig. 4. Results of Testing the Proposed Representation on Natural Imagery. The input imagery and general format of the display are the same as in Fig. 3. Four additional components of the representation are now shown to incorporate information in the y spatial dimension. The overall pattern of results are consistent with predictions.

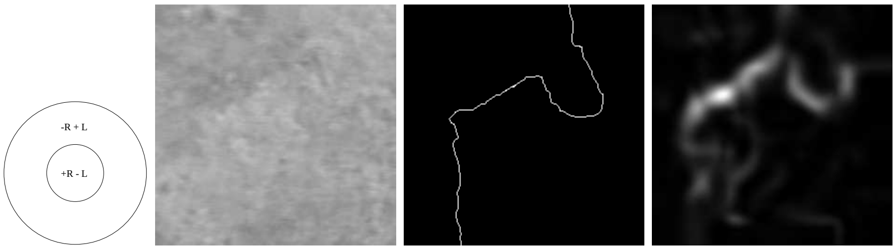

Fig. 5. Motion Boundary Detection. Left to right: A schematic of a double opponent motion operator for motion boundary detection. An aerial image of a tree canopy moving against undergrowth with relative motion due to camera movement. The hand marked outline of the motion boundary. The magnitude of the boundary signal. The result accurately localizes the edge of the tree against the background.