The Applicability of Spatiotemporal Oriented Energy Features to Region Tracking

Kevin J. Cannons and Richard P. Wildes

Abstract—This paper proposes the novel application of an uncommonly rich feature representation to the domain of visual tracking. The proposed representation for tracking models both the spatial structure and dynamics of a target in a unified fashion, while simultaneously offering robustness to illumination variations. Specifically, the proposed feature is derived from spatiotemporal energy measurements that are computed by filtering in $3D$, (x, y, t), image spacetime. These spatiotemporal energy measurements capture the underlying local spacetime orientation structure of the target across multiple scales. The breadth of applicability of these features within the field of visual tracking is demonstrated by their instantiation within three disparate tracking paradigms that are representative of the various basic types of region trackers in the field. Instantiation within these three tracking paradigms requires that the raw oriented energy measurements be post-processed using different methodologies that range from histogram accumulation to the identity transform. Qualitative and quantitative empirical evaluation on a challenging suite of videos demonstrates the strength and applicability of the proposed representation to tracking, as it outperforms other commonly-used features across all tracking paradigms. Moreover, it is shown that overall high tracking accuracy can be obtained with this proposed representation, as spatiotemporal oriented energy instantiations are shown to outperform several recent, state-of-the-art trackers.

Index Terms—Visual tracking, feature representations, motion analysis, spatiotemporal orientation, visual spacetime

1 Introduction

1.1 Motivation

Visual tracking is a core area of computer vision, with both theoretical and practical significance. Even given this strong motivation, to date a general purpose visual tracker that operates robustly across all real-world settings has not emerged. One key challenge for visual trackers is illumination effects. Under the use of many popular representations (e.g., colour), the features' appearance changes drastically depending on the lighting conditions. A second challenge for visual trackers is clutter. As the amount of scene clutter increases, so to does the chance that the tracker will be distracted away from the true target by other "interesting" scene objects (i.e., objects with similar feature characteristics). Finally, trackers often experience errors when the target exhibits sudden changes in appearance or velocity that violate the underlying assumptions of the system's models.

In this work, it is proposed that the choice of representation is key to meeting the above challenges. A representation that is invariant to illumination changes will be better able to track through significant lighting effects. A feature set that provides a rich characterization will be less likely to confound the true target with other scene objects. Finally, a rich representation allows for greater tracker resilience to sudden changes in appearance or velocity. This resilience is attained because the tracker can rely on stable, more consistent components of the representation as one or more components experience a fast change. In the current paper, a spatiotemporal oriented energy (SOE) representation is shown to address the above requirements and is applied for the first time to the domain of visual tracking. This representation uniformly captures both the spatial and dynamic properties of the target for a rich characterization, with robustness to illumination and amenability to on-line updating.

• *K. Cannons and R. Wildes are with the Department of Computer Science and Engineering and Centre for Vision Research (CVR), York University, Toronto, Canada.*
E-mail: kcannons, wildes@cse.yorku.ca

1.2 Related research

Visual trackers can be coarsely divided into three general categories: (i) discrete feature trackers (ii) contour trackers, and (iii) region trackers [1], [2]. Since the present contribution falls into the region tracker category, only the most relevant works in this class will be reviewed. Some region trackers isolate moving regions of interest by performing background subtraction and data association between the detected foreground "blobs" [3], [4], [5], [6], [7], [8]. A limitation of the blob tracking approach is that the systems tend to assume stationary cameras and rely on background subtraction techniques, which are often noisy.

Another category of region trackers that has seen significant recent research leverages detectors that are trained using offline [9], [10] or online [11], [12], [13], [14], [15], [16], [17] machine learning techniques.

Digital Object Indentifier 10.1109/TPAMI.2013.233

Offline methods tend to make use of human detectors [18] to identify potential target locations, which are subsequently linked via data association [9], [10]. These offline systems have the limitation that they can only track one type of target (the object upon which the detector was trained). Systems employing online target detection methods learn the properties of the target in the first frame and subsequently update the internal target template thereafter [11], [12], [16], [14].

Region trackers have also been considered that seek efficiency through more compact target representations. Such systems collapse spatial information across the target support and use histogram representations of the target during tracking [19], [20], [21], [22], [23], [24]. Earlier instantiations of this approach lacked discriminability owing to collapse of spatial layout information across target support [19], which led to subsequent refinements that preserved various amounts of target layout [25], [20], [21], [22], [24].

A final type of region tracker retains spatial organization within the tracked area by using (dense) pixelwise feature measurements [26], [27], [28], [29], [30], [31], [32], [33], [13]. Building on early image alignment work [26], pointwise warp (PW) tracking subsequently was refined to include more advanced motion models [34], robust error metrics [35], and more sophisticated appearance model updating mechanisms [30], [31].

A key underlying dimension that distinguishes region trackers is the degree to which they aggregate spatial information across target support: Some completely collapse spatial information (e.g., mean shift [19]); others consider a coarse measure of spatial layout (e.g., FragTrack [22]); still others maintain complete target spatial layout (e.g., [35]). Depending on the tracking task at hand, the amount of spatial layout information that should be used for best accuracy may vary and research has considered automatically adapting tracker operation along this dimension, e.g., [36]. Since the amount of spatial layout retained is a critical dimension of tracker design, the feature set proposed in the current paper will be evaluated at three different points along this spectrum.

A common issue that connects region trackers of all types is the base features or measurements that are used. The most commonly used features in region tracking are pixel intensity (e.g., [26], [37], [22]) and color (e.g., [38], [19], [39]). Color and intensity are notoriously sensitive to illumination changes, which has prompted the use of various color spaces with some limited improvement, e.g., [20], [21].

Another commonly-used feature representation is the image gradient, e.g., [40], [21]. An advantage of image gradients is that they often remain more consistent throughout illumination changes in comparison to color or intensity; however, they also produce spurious responses from heavily textured materials and in clutter. Another similar representation considers the output of orientation selective filters [30]. In a similar fashion to color features (e.g., RGB channels), this representation consists of multiple measurements or "channels" at each location.

Recovered motion is another cue that can be used as a feature for tracking (e.g., [41], [42], [43], [15]). Motion-based cues can be beneficial in camouflage situations where a target is nearly indistinguishable from the background when considering its appearance in isolation, but can be identified clearly from its dynamics. However, motion cues may not be sufficiently discriminative in isolation when the targets are prone to velocity changes or exhibit similar motion patterns to other scene objects.

Although a range of feature representations has been considered, each of these commonly-used representations is prone to failure in certain situations. In an attempt to leverage the complementary nature of various features, numerous trackers have considered combining cues [44], [41], [45], [46], [47], [48], [49], [33], [15]. One popular approach is to combine color with edges [44], [47], [33]. Other approaches have considered joint feature-spatial spaces that combine an explicit positional feature with other cues (e.g., [45], [46]). The combination of appearance cues with recovered motion has also been fruitful (e.g., [41], [48], [49], [15]). Although feature combinations have demonstrated improved tracking performance, there are drawbacks of this approach: (i) Effective cue integration remains an open problem. (ii) Feature extraction requires additional processing. (iii) The designer must ensure that a sufficient number and variety of features are available for a given tracking task.

A research tack that arguably requires greater attention is that of deriving new feature sets for visual tracking. In particular, limited attention has been given to the development of uniformly derived feature representations for tracking that encompass both spatial and temporal domains. Potential benefits of an integrated approach include the ability to combine static and dynamic target information in a natural fashion as well as simplicity of design and implementation. In response to this observation, this paper proposes a feature representation, based on spatiotemporal oriented energies, that has not been used previously in the context of visual tracking. These energies provide a particularly detailed target description, with robustness to illumination, that captures both appearance and dynamics in an integrated fashion. It should be stressed that similar spatiotemporal oriented energy features have been applied with success to other visual information processing tasks, e.g., motion estimation [50], [51], action spotting [52], [53], dynamic texture analysis [54], qualitative descriptors [55], and spacetime stereo [56], *but never before to visual tracking*.

1.3 Contributions

In light of previous research, the main contributions of the present paper are as follows.

- A spatiotemporal oriented energy (SOE) representation, similar in spirit to representations that have been demonstrated with success in other areas of computer vision, is uniquely applied to visual tracking. This representation offers an extremely rich target representation, capturing both appearance and motion cues, while simultaneously providing significant robustness to illumination effects.
- Theoretical methods are developed for instantiating the SOE features in three disparate tracking architectures. Specifically, theoretical techniques are derived showing that the SOE features can be used to construct a histogram representation for use in mean shift tracking, to represent target subregions with a number of histograms in the FragTrack framework, and as pointwise measurements in the warping paradigm.
- The discriminative power of the proposed SOE representation is demonstrated via a direct comparison against other commonly-used features. Four feature types in three different tracking architectures are considered, with a total of eleven trackers used for direct comparison of SOEs vs. other features. *The SOE representation offers overall superior tracking accuracy when compared to alternative features using the same tracking paradigm*.
- In addition to the eleven-way feature set comparison, five additional state-of-the-art tracking systems are evaluated. When the entire set of sixteen algorithms are evaluated on a suite of nine challenging videos, *one of the proposed SOE-based trackers yields best overall tracking accuracy*. The other two trackers based on SOEs are among the top five best overall systems and show superior performance to recent strong trackers.

Preliminary versions of this research have been reported previously [23], [57].

2 Technical Approach

2.1 Spatiotemporal oriented energies (SOEs)

Video sequences induce very different orientation patterns in image spacetime depending on their contents. For instance, a textured, stationary object yields a much different orientation signature than if the very same object were undergoing translational motion. An efficient framework for analyzing spatiotemporal information can be realized through the use of 3D, (x, y, t), oriented energies [58]. These energies are derived from the filter responses of orientation selective bandpass filters that are applied to the spatiotemporal volume representation of a video. A chief attribute of an oriented energy representation is its ability to encompass both spatial and dynamic aspects of visual spacetime, strictly through the analysis of 3D orientation. Consideration of spatial patterns (e.g., textures) is performed when the filters are applied within the image plane. Dynamic attributes of the scene (e.g., velocity and flicker) are analyzed by filtering at orientations that extend into the temporal dimension.

The desired oriented energies are realized using broadly tuned 3D Gaussian second derivative filters, $G_2(\theta, \gamma)$, and their Hilbert transforms, $H_2(\theta, \gamma)$, where θ specifies the 3D direction of the filter axis of symmetry, and γ indicates the scale. Thus, the oriented energies provide a local decomposition in terms of angular, θ, and radial, γ, frequency, where the former captures the local directionality of image structure and the latter encompasses the local granularity. In particular, to attain an initial measure of energy, the filter responses are pointwise rectified (squared) and summed according to

$$E(\mathbf{x}; \theta, \gamma) = [G_2(\theta, \gamma) * I(\mathbf{x})]^2 + [H_2(\theta, \gamma) * I(\mathbf{x})]^2 , \quad (1)$$

where $\mathbf{x} = (x, y, t)$ are spatiotemporal image coordinates, I is an image, $*$ denotes convolution, and care should be taken to normalize the filters to ensure that their energy across scales is constant [59].

The initial definition of local energy measurements, (1), is dependent on image contrast (i.e., it will increase monotonically with contrast). To obtain a purer measure of the relative contribution of orientations irrespective of image contrast, pointwise normalization is performed,

$$\hat{E}(\mathbf{x}; \theta, \gamma) = \frac{E(\mathbf{x}; \theta, \gamma)}{\sum_{\theta_i} \sum_{\gamma_i} E(\mathbf{x}; \theta_i, \gamma_i) + \epsilon} , \quad (2)$$

where ϵ is a constant introduced as a noise floor and to avoid numerical instabilities when the overall energy content is small. Additionally, θ_i and γ_i consider all orientations and scales, respectively.

For illustrative purposes, Fig. 1 displays a subset of the energies that are computed for a single frame of a traffic sequence. The scene involves a white car moving to the left near the center of a traffic intersection. Notice how the energy channels that are tuned for leftward motion are very effective at distinguishing this car from the static background. Consideration of the channels tuned for horizontal structure show how they capture the overall orientation structure of the white car. In contrast, while the channels tuned for vertical textures capture the outline of the crosswalks, they show little response to the car, as it is largely devoid of vertical structure at the scales considered. Finally, note how the energies become more diffuse and capture more gross structure at the coarser scale.

From the theoretical development as well as the illustrative example, it appears that spatiotemporal oriented energies are well-suited to form the feature representation in visual tracking applications for four significant reasons:

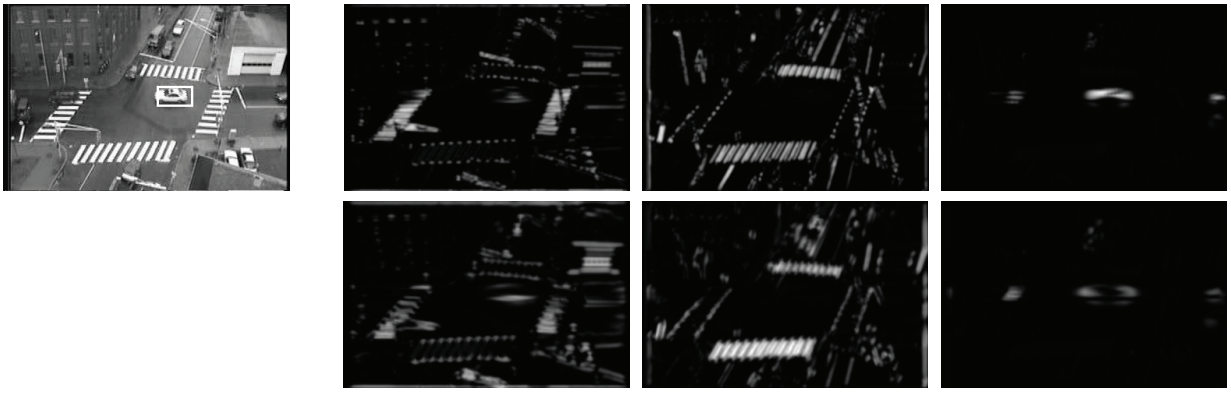

Fig. 1. Illustrative example showing a sample of the spatiotemporal oriented energy representation captured for a single video frame. (Top Left) Frame of a MERL traffic video sequence [60]. (Top and Bottom Rows, Columns Two to Four) Selected spatiotemporal oriented energy channels corresponding to the input video frame. Finer and coarser scales are shown in top and bottom rows, resp. From left to right, the energy channels roughly correspond to horizontal structure, vertical structure, and leftward motion.

- A rich description of the target is attained due to the fact that oriented energies encompass both target appearance and dynamics. This richness will be shown to allow for trackers that are more robust to clutter both in the form of background static structures and other moving targets in the scene. Since the representation contains both spatial and dynamic attributes, targets and distractors can be differentiated based on velocities if their appearances are similar, or vice versa.
- The oriented energies are robust to illumination changes. By construction, the proposed representation provides invariance to both additive and multiplicative intensity changes. Invariance to additive biases is achieved through the process of bandpass filtering, (1); whereas multiplicative biases are removed via normalization, (2).
- The energies can be computed at multiple scales, allowing for a multiscale analysis of the target attributes. Finer scales provide information regarding motion of individual target parts (e.g., limbs) and detailed spatial textures (e.g., facial expressions, clothing logos). Complementary coarser scales provide information regarding the overall target velocity and its gross shape.
- The oriented energies are efficiently computed via linear and pointwise non-linear operations [61], with amenability to real-time realizations on GPUs [62].

To demonstrate the widespread applicability of SOE features to visual tracking, a representative set of trackers from the huge set of all region trackers is considered. As noted in the discussion of related work, a key dimension along which region trackers vary is the amount of spatial layout information they employ. Correspondingly, in the remainder of this section three representative trackers along this dimension are selected for instantiation with SOE features: mean shift tracking [19] (zero spatial arrangement information), fragment tracking [22] (limited spatial arrangement information), and pointwise warp tracking [35] (complete spatial arrangement information).

2.2 SOE mean shift tracking

The energies as defined in (2) exhibit broad response patterns, particularly at coarser scales. This property is illustrated in Fig. 1. The overly diffuse responses emerge due to the particular filters that are employed as well as the downsampling/upsampling that is utilized in pyramid processing. However, given that the tracking problem is being considered, the goal is to locate the target's position as precisely as possible. Coarse energies remain important because they provide information regarding the target's gross shape and motion, but their localization must be improved for accurate tracking. To that end, a set of weights are applied to the normalized energies, (2), according to

$$\hat{E}^{*}\left(\mathbf{x};\theta,\gamma\right)=\hat{E}\left(\mathbf{x};\theta,\gamma\right)z\left(\mathbf{x};\theta\right), \tag{3}$$

where z are pixelwise weighting factors for a particular orientation channel, θ. The weighting factors for a specific orientation are computed by integrating the energies across all scales and applying a threshold, Z_θ, according to

$$z\left(\mathbf{x};\theta\right)=\begin{cases}1, & \text{if } \sum_{\gamma_i}\hat{E}\left(\mathbf{x};\theta,\gamma_i\right)>Z_\theta\\ 0, & \text{otherwise}\end{cases}. \tag{4}$$

Note that there is a separate threshold, Z_θ, applied to each orientation θ, but these thresholds are uniformly and automatically derived from the image data. Specifically, in this work, the thresholds are set based on the average energy in each orientation channel, θ. When computing the weights, z, summing across scales allows the better localized fine scales to sharpen the coarse scales, while the coarse scales

help to smooth the responses of the fine scales. Furthermore, by calculating weights separately for each orientation, prejudice toward any particular type of oriented structure (e.g., static vs. dynamic) is avoided.

Interestingly, while application of these weights was found to improve the performance of SOE mean shift tracking, it had little impact on fragment and pointwise warp instantiations, so is not included therein. Apparently, the loss of target spatial layout information incurred through holistic histogram accumulation in mean shift tracking demands increased precision in energy localization relative to the alternative paradigms that maintain increased target layout.

As the mean shift tracking paradigm is being considered in this section, the spatial information is collapsed to represent the target as a histogram. Unlike the traditional mean shift tracker that employs color histograms [19], here a spatiotemporal oriented energy histogram is constructed. Specifically, each bin in the histogram corresponds to the weighted energy content of the target at a particular scale and orientation. The template histogram that defines the target in the first frame is given by

$$q_u = C \sum_i k\left(\|\mathbf{x}_i\|^2\right) \hat{E}^*\left(\mathbf{x}_i, t_0; \phi_u\right) \quad , \tag{5}$$

where k is the profile of the tracking kernel, C is a normalization constant to ensure the histogram sums to unity, $\mathbf{x}_i = (x, y)$ is a target pixel at some temporal instant, i ranges so that $\mathbf{x}_i$ covers the template support, and ϕ_u is the scale and orientation combination which corresponds to bin u of the histogram.

To evaluate target candidates in a current frame, candidate histograms are defined as

$$p_u\left(\mathbf{y}\right) = C_h \sum_i k\left(\left|\left|\frac{\mathbf{y} - \mathbf{x}_i}{h}\right|\right|^2\right) \hat{E}^*\left(\mathbf{x}_i, t; \phi_u\right) \quad , \tag{6}$$

where $\mathbf{y}$ is the center of the target candidate's tracking window, h is the bandwidth of the tracking kernel and i ranges so that $\mathbf{x}_i$ covers the candidate support.

A sample energy histogram for the target region shown in Fig. 1 (represented by the white box) is shown in Fig. 2. The bin corresponding most closely to leftward motion at the finest scale (bin 5) has by far the most energy. The next two high energy counts are found in bins 2 and 9 which are tuned to combinations of dynamic and static structure, with an emphasis on leftward motion and spatial orientation similar to that of the target. The overall horizontal structure of the car is captured by the energy in bins 1 and 4. In contrast, bins 3 and 6, which roughly represent static, vertical structure, do not have strong responses, given the nature of the car target. The histogram also shows that the SOEs for the highest frequency structures have the strongest response, as the target is small and dominated by relatively finer scale structure.

With template, $\mathbf{q}$, and candidate, $\mathbf{p}(\mathbf{y})$ histograms defined in terms of SOE features, (5) and (6), resp., the mean shift tracking instantiation otherwise follows the original approach [19]. Histograms with m bins are compared using the Bhattacharyya coefficient [63],

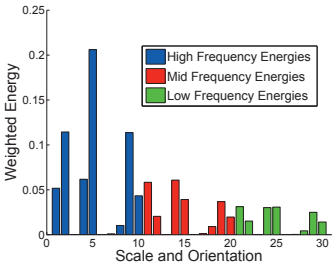

Fig. 2. SOE histogram for the target region in Fig. 1.

$$\rho\left[\mathbf{p}\left(\mathbf{y}\right), \mathbf{q}\right] = \sum_{u=1}^{m} \sqrt{p_u\left(\mathbf{y}\right) q_u} \quad , \tag{7}$$

which is maximized with respect to target position, $\mathbf{y}$ using mean shift iterations starting from the target's position in the previous frame.

2.3 SOE FragTrack

In a similar manner to the mean shift tracker, the original FragTrack system operates upon a feature histogram representation of the target [22]. However, rather than collapsing all information regarding the target's spatial layout, FragTrack represents the target using a number of histograms, extracted from different rectangular regions within the target support.

To instantiate SOE features within the FragTrack framework, for each rectangular patch defined in the target support, an energy histogram is constructed in a similar manner to those proposed for the SOE mean shift tracker. Specifically, a histogram bin for the j^{th} fragment in the template is defined according to

$$\bar{q}_u^j = \bar{C} \sum_i \hat{E}\left(\mathbf{x}_i, t_0; \phi_u\right), \tag{8}$$

where $\bar{C}$ is a normalization constant to ensure the histogram sums to unity and $\mathbf{x}_i = (x, y)$ is a single target pixel at some initial temporal instant, t_0. Additionally, i ranges so that $\mathbf{x}_i$ covers the template support of the j^{th} fragment and ϕ_u is the scale and orientation combination that corresponds to bin u of the histogram. Similarly, a histogram bin for the j^{th} fragment within a candidate at time, t, is defined as

$$\bar{p}_u^j = \bar{C} \sum_i \hat{E}\left(\mathbf{x}_i, t; \phi_u\right). \tag{9}$$

Note that for the FragTrack equations, overbars are used to differentiate between the analogous histogram parameters that are used for mean shift tracking. Although the energy histograms constructed for the fragments share similarities with the holistic ones used in the SOE mean shift framework, an important difference is the lack of a spatial weighting kernel.

With the SOEs instantiated as fragment histograms, FragTrack tracking can follow the standard procedure [22]. To compare template to candidates in a current frame, the Bhattacharrya coefficient, (7), is employed once again. To combine coefficient scores across j fragments, the scores for each candidate in the search region, $\rho\left[\bar{\mathbf{p}}^{j}\left(\mathbf{y}\right),\bar{\mathbf{q}}^{j}\right]$, are sorted from best to worst. Comparisons between candidates are made using the fragment with the s^{th} best score. The motivation is that s is selected such that it is the maximum number of fragments that are expected to provide inlier measurements and thereby provides robustness to partial target occlusions. Making these ideas more precise, let j^{*} indicate the fragment with the s^{th} best score for a particular candidate position, $\mathbf{y}$. The overall score of a candidate is evaluated as

$$\Omega\left(\mathbf{y}\right)=\rho\left[\bar{\mathbf{p}}^{j^{*}}\left(\mathbf{y}\right),\bar{\mathbf{q}}^{j^{*}}\right]. \tag{10}$$

To locate the target within a current frame, the similarity equation, (10), is optimized over $\mathbf{y}$ via exhaustive search within a radius, D, of the estimated target position in the previous frame. Integral images are used for histogram construction to enable efficient search.

2.4 SOE pointwise warp tracking

Unlike the mean shift SOE tracker of Section 2.2 and the SOE FragTrack system of Section 2.3 that perform some degree of spatial accumulation over the energy measurements, warping trackers employ pointwise feature measurements; therefore, complete information regarding the target's spatial configuration is retained. In this framework, the normalized, pointwise energy features, (2), of Section 2.1 are incorporated directly to define the target. In particular, the first frame template is defined as

$$Q\left(\mathbf{x},\phi_{u}\right)=\hat{E}\left(\mathbf{x},t_{0};\phi_{u}\right), \tag{11}$$

for energies measured at some start time, t_0, and spatial support, $\mathbf{x}=(x,y)$, over some suitably specified region. Additionally, $\phi_u=(\theta_u,\gamma_u)$ is the orientation and scale combination that corresponds to a particular channel of data, u. Thus, the template is indexed spatially by position, (x,y), and at each position it provides a set of $\theta\times\gamma$ energy measurements that indicate the relative presence or absence of spacetime orientations. The candidate feature images are defined similarly according to

$$P\left(\mathbf{x},\phi_{u},t\right)=\hat{E}\left(\mathbf{x},t;\phi_{u}\right), \tag{12}$$

where t indicates the current time instant.

Note that the target representation, (12), is in contrast to standard template tracking-based systems that typically only utilize a single channel of intensity features during estimation [26], [34], [35]. Further, even previous approaches that have considered multiple measurements/pixel make use of only spatially derived features (e.g., [30]), which will be shown in Sec. 3 to significantly limit performance in comparison to the proposed pointwise warp tracker.

Tracking using a PW approach consists of matching the template, Q, to the current frame of the sequence so as to estimate and compensate for the interframe motion of the target. In the present approach, both the template, Q, and a candidate from the current image frame, P, are represented in terms of oriented energy measurements, (2). An affine motion model is used to capture target interframe motion, as applicable when the target depth variation is small relative to the camera-to-target distance [64], [27], [31]. The affine motion model is defined explicitly as $\mathbf{u}\left(x,y;\mathbf{a}\right)=\left(a_0+a_1x+a_2y,a_3+a_4x+a_5y\right)^{\top}$ where $\mathbf{a}=\left(a_0,a_1,\ldots a_5\right)^{\top}$ are the motion parameters and (x,y) are pixel coordinates.

The affine parameters, $\mathbf{a}$, are estimated by minimizing an error function that is derived from the optical flow constraint equation (OFCE). Since the target representation spans not just a single spatial image plane, but multiple feature channels (orientations and scales) of SOEs, error minimization is performed across the target spatial support and over all feature channels. To measure deviation from the OFCE, a robust error metric, $\psi(\eta,\sigma)$, is utilized [35], which is beneficial for occlusions, imprecise target delineations that include background pixels, and target motion that deviates from the affine motion model (e.g., non-rigid, articulated motion). The resulting error to be minimized with respect to $\mathbf{a}$ is

$$\sum_{\mathbf{x}}\sum_{\theta}\sum_{\gamma}\psi\left[\nabla^{\top}P\left(\mathbf{x};\theta,\gamma\right)\mathbf{u}\left(\mathbf{a}\right)+P_{t}\left(\mathbf{x};\theta,\gamma\right),\sigma\right], \tag{13}$$

where $\nabla^{\top}P=\left(P_x,P_y\right)$ are the first-order spatial derivatives of the image energy measurements in the current frame and $P_t=P-Q$ denotes the first order temporal derivative found by computing the difference between the candidate and aligned template. In the present implementation, the Geman-McClure error metric [65] is utilized with σ, the robust metric width, as suggested in [35]. Minimization to yield the motion estimate, (13), is performed using a gradient descent procedure [35]. To increase the capture range of the tracker, the minimization process is performed in a coarse-to-fine fashion [34], [35].

2.5 Template and scale updates

When tracking an object through a long video sequence, it is common that its characteristics will change. For the proposed SOE trackers, template adaptation is necessary to ensure that changes in target appearance (e.g., target rotation, addition/removal of clothing accessories, changing facial expression) and dynamics (e.g., speeding up, slowing down, changing direction) are accurately represented by the current template. To combat such changes,

the proposed SOE trackers include straightforward autoregressive type template update mechanisms. The template update equation for the SOE mean shift tracker is defined as

$$\mathbf{q}^{t+1} = \alpha_{\mathrm{MS}} \pi \mathbf{q}^{t} + (1 - \alpha_{\mathrm{MS}})(1 - \pi)\,\mathbf{p}^{*} \quad , \tag{14}$$

where α_{MS} is a weighting factor to control the speed of the updates, $\mathbf{q}^t$ is the template at frame t, and $\pi = \rho[\mathbf{p}(\mathbf{y}^*), \mathbf{q}^t]$ is the Bhattacharyya coefficient between the current template and the optimal candidate found in the t^{th} frame at $\mathbf{y}^*$. Additionally, $\mathbf{p}^* = \mathbf{p}(\mathbf{y}^*)$ denotes the optimal candidate histogram in the current frame. Following each application of (14), the resulting template is renormalized and thereby remains consistent with the overall formulation.

For the SOE FragTrack and PW systems, a similar template update is utilized

$$\mathbf{Q}_{\mathrm{Alg}}^{t+1} = \alpha_{\mathrm{Alg}} \mathbf{Q}_{\mathrm{Alg}}^{t} + (1 - \alpha_{\mathrm{Alg}})\,\mathbf{P}_{\mathrm{Alg}}^{*} \quad , \tag{15}$$

where Alg indicates the tracking algorithm (i.e., $\mathrm{Alg} = \{\mathrm{FragTrack}, \mathrm{PW}\}$. Additionally, $\mathbf{Q}$ denotes the template data structure for the two trackers, while $\mathbf{P}^*$ is the optimal candidate region identified in the current frame. Equations (14) and (15) differ slightly in that the mean shift update includes additional weighting terms based on the Bhattacharyya match score, π. A comparable term could be included in (15) but empirical evaluation revealed this was unnecessary.

The size of a target may change during a video sequence as well. The SOE mean shift and FragTrack scale update mechanisms follow those originally proposed for those paradigms [19] and [22], resp. Size changes for the SOE PW system are handled by the scaling afforded by the affine tracking transformation.

Interestingly, the simple template update mechanisms used here, even though not representative of the state-of-the-art in adaptation [30], [66], [67], [31], [12], allow the SOE trackers to achieve competitive results, owing to their feature representation strength. In particular, the richness of the representation, capturing both spatial and dynamic target properties, allows for the relatively stable components of the representation to keep the tracker on target during changes, while the altered components adapt using the update mechanisms provided.

3 Empirical Evaluation

3.1 Data Sets and Tracker Settings

Extensive empirical evaluation has been conducted, comparing the performance of sixteen distinct trackers on a suite of nine challenging and publicly available video sequences. The suite of video sequences span the outstanding challenges of visual tracking, including drastic illumination changes, small targets, scene objects with similar visual appearance to the target, appearance changes, and occlusions. Two of the sequences also involve non-stationary cameras.

TABLE 1
Videos used in empirical evaluation.

Video	Description
Occluded Face 2 [12]	Facial target. In plane target rotation. Cluttered background. Appearance change via addition of hat. Significant occlusion by book and hat.
Tiger 2 [12]	Hand-held toy animal target. Small target with fast, erratic, and articulated motion. Cluttered background/foreground. Occlusion as target moves amongst leaves.
Sylvester [31]	Hand-held stuffed animal target. Fast erratic motion, including out-of-plane rotation/shear. Illumination change across trajectory.
Illumination [23]	Human target with significant non-rigid deformations. Drastic lighting changes between lit and unlit areas. Cluttered background. Video available in color.
PETS [68]	Extremely small cyclist target with indistinct color. Cluttered background. Partial occlusion behind pedestrian. Video available in color.
Ming [69]	Facial target. In-plane and out-of-plane rotation. Significant illumination and scale changes.
Woman [22]	Deformable human target with panning and zooming camera. Extended and significant partial occlusions. Video available in color.
Car11 [31]	Car rear target with pursuit camera. Night time scene with harsh lighting and low contrast. Video available in color.
Pop Machines [57]	Multiple similar appearing targets with crossing trajectories. Low quality surveillance video. Harsh lighting. Full occlusion from pillar.

Of the nine videos, four are available both in color and greyscale. The video sequences that are considered in this evaluation are summarized in Table 1. Manually labelled, ground truth tracking windows and their corresponding videos are available for download at http://www.cse.yorku.ca/vision/research/visual-tracking/.

The tracking algorithms used for comparison include the three proposed in Section 2 as well as thirteen additional benchmark systems. These benchmark systems serve two purposes. Eight of the benchmark trackers are included to evaluate the performance of the proposed SOE representation against commonly used feature sets (i.e., intensity, color, and purely spatial oriented energies). Further, five recent and state-of-the-art trackers are included to measure absolute performance of the proposed trackers. Table 2 describes the sixteen trackers in more detail.

The systems that employed spatiotemporal oriented energies (i.e., **PW-SOE**, **Frag-SOE**, and **MS-SOE**), all utilized the same underlying features. Specifically, SOEs were computed at 10 orientations, as they span the space of 3D orientations for the highest order filters that were used (i.e., H_2). The particular orientations selected were the normals to the faces of an icosahedron, as they evenly sample the sphere. For the trackers that operated upon purely spatial oriented energies (i.e., **PW-OE**, **Frag-OE**, and **MS-OE**), the features were computed at four orientations ($0°$, $45°$, $90°$, and $135°$), so as to span the space of 2D orientations for the highest order filters. Both

TABLE 2
Trackers used in empirical evaluation.

Tracker	Description
PW-SOE	Pointwise warp tracker with spatiotemporal oriented energy features (Sec. 2.4).
PW-OE	Pointwise warp tracker with purely spatial oriented energy features.
PW-Int	Pointwise warp tracker with intensity features.
Frag-SOE	FragTrack system with spatiotemporal oriented energy features (Sec. 2.3).
Frag-OE	FragTrack system with purely spatial oriented energy features.
Frag-Int	FragTrack system with intensity features [22].
Frag-Color	FragTrack system with color features [22].
MS-SOE	Mean shift tracker with spatiotemporal oriented energy features (Sec. 2.2).
MS-OE	Mean shift tracker with purely spatial oriented energy features.
MS-Int	Mean shift tracker with intensity features [19].
MS-Color	Mean shift tracker with color features [19].
IVT	State-of-the-art incremental visual tracker [31].
MIL	State-of-the-art multiple instance learning tracker [12].
A-BHMC	State-of-the-art adaptive basin hopping Monte Carlo based tracker [32]. This tracker operates on color features.
VTD-Int	State-of-the-art visual tracking decomposition system using intensity and edge features [33].
VTD-Color	State-of-the-art visual tracking decomposition system using hue, saturation, intensity, and edge features [33].

spatial and spatiotemporal energies were computed at a single scale, corresponding to direct application of the oriented filters to the input imagery. The intensity-based trackers (i.e., **PW-Int**, **Frag-Int**, and **MS-Int**) operated directly on the greyscale video pixels. The color-based implementations (i.e., **Frag-Color** and **MS-Color**) operated on the RGB color pixels and formed RGB histograms. No color instantiation of the **PW** type tracker is included, as use of color features in that framework is uncommon.

Parameter settings for all systems were determined empirically, such that tracking accuracy was optimized. For the pointwise warp trackers, motion estimation was performed using coarse-to-fine processing operating over four levels of a Gaussian pyramid built on top of the feature measurements. Template updates were performed with a rate of $\alpha_{\mathrm{PW}} = 0.999$.

The FragTrack systems employed a similar fragment topology to that suggested in [22]. Template updates used the setting $\alpha_{\mathrm{Frag}} = 0.99$. Additionally, a search radius of $D = 45$ pixels from the previous target position was utilized and the $s = 50^{\mathrm{th}}$ percent quantile was used when selecting the fragment for use in candidate comparisons, (10).

For the mean shift trackers, the system using SOEs employed energy thresholds, Z_θ, of $2.75\times$ the mean energy for each orientation channel. With spatial oriented energies, best performance was found using $Z_\theta = 0$; whereas this parameter is not relevant to **MS-Int** since it operates on scalar features. Template updates were set according to $\alpha_{\mathrm{MS}} = 0.85$. Finally, the Epanechnikov kernel was used and the maximum number of mean shift iterations was set to twenty.

With regards to the recent state-of-the-art systems, **IVT** [31], **MILTrack** [12], **A-BHMC** [32], and **VTD-Int**/**VTD-Color** [33] parameters were assigned according to the original authors' suggestions or with values that were experimentally validated as providing superior performance. All four of these state-of-the-art trackers have been used extensively in empirical evaluations, with particular emphasis on **IVT**, **MILTrack**, and **VTD-Int** in most recent evaluations (e.g., [70], [36], [17]). It should also be noted that this evaluation considers implementations comparable to the original mean shift [19] (**MS-Int**) and FragTrack [22] (**Frag-Int**), which also are considered strong trackers, e.g. [15], [70]. For all trackers, parameters were held constant throughout all experiments. Moreover, all trackers were provided with identical first-frame initializations, which were annotated by hand.

3.2 Qualitative results

Figure 3 displays qualitative tracking results for all trackers that employed SOE features as well as the three state-of-the-art systems that consider greyscale features (i.e., **IVT**, **MIL**, and **VTD-Int**). Results for the ten remaining trackers are suppressed here due to space constraints; however, they are presented in Section 3.3 when quantitative performance results are analyzed. As the figure shows, for *Occluded Face 2*, **PW-SOE**, **IVT**, and **VTD-Int** provide comparable qualitative results; whereas **MIL** becomes more poorly localized during the later stages of the video. Although **VTD-Int** provides a well-centered and well-scaled track of the target throughout, this system does not estimate rotation; thus, it does not accurately capture the rotations as the target tilt's his head in Frame 424. **Frag-SOE**, **MS-SOE**, and **MIL** share this limitation as they too do not estimate rotational target motion. Finally, the collapsing of spatial arrangement information in conjunction with a loose initial target window limits the performance of both **Frag-SOE** and **MS-SOE**. With such impoverished representations of spatial layout information, clutter in the background is easily confounded with the true target.

In *Tiger 2*, **MIL** and **Frag-SOE** provide the best tracking accuracy throughout. Both trackers make use of a "spotting approach", where an exhaustive search is performed within a search radius, which is better-suited for following targets moving at extreme velocities. In contrast, **PW-SOE** struggles somewhat relative to **MIL** and **Frag-SOE** because the small target combined with rapid motion makes it difficult for the employed coarse-to-fine, gradient-based motion estimator to obtain accurate updates. The result for **PW-SOE** is that it lags behind during the fastest motions; although, it "catches-up" throughout. The **VTD-Int** system initially offers competitive performance to

MIL and **Frag-SOE**, but it loses the target at Frame 280 and never reacquires it. Further, the reduced discriminative power offered by the histogram representation used in **MS-SOE** again limits its performance. Thus, the mean shift system is often attracted to background clutter (e.g., Frame 74), but it nonetheless tends to lock back on to the target eventually (e.g., Frame 222). For this video, **IVT** offers worst performance, as it falls off target early in the sequence (approximately Frame 100) and does not reacquire the target.

In *Sylvester*, both **Frag-SOE** and **IVT** experience complete tracking failures early in the sequence. For **Frag-SOE**, an abrupt change in velocity accompanied by an appearance change (resulting from a slight target rotation) was sufficient to disrupt tracking. On the other hand, **IVT** lost the target slightly later, when it suddenly rotates toward the camera (rapid appearance change). **MIL** follows the target throughout the entire sequence, but at times the lighting and appearance changes (due to out-of-plane rotations) move the tracking window partially off-target. **MS-SOE** also tracks the target throughout the sequence, but allows its tracking window to grow gradually too large due to a relatively unstructured background and no notion of target spatial organization. The **VTD-Int** system provides an accurate target track throughout the majority of the video but experiences a complete tracking failure at Frame 1095 when the target undergoes a dramatic out-of-plane rotation from which it never recovers. Finally, **PW-SOE** performs best due to the robustness of its pointwise features to illumination changes and their ability to capitalize on motion information when appearance varies rapidly.

For the *Illumination* sequence, all three SOE-based systems display accurate and comparable tracking accuracy throughout, due to the robustness of this representation to illumination changes, as discussed in Section 2.1. The performance of **IVT**, **MIL**, and **VTD-Int** suffers in this sequence, due to the extreme lighting effects and lack of robustness to such challenges in the representations that are utilized. Specifically, **IVT** loses the target at Frame 13, when it emerges from the shadows; while **MIL** and **VTD-Int** hold on longer, they also fail when the target reenters the shadows. It appears that the Haar-like features used in **MIL** and the inclusion of edge features in **VTD-Int** lead to some limited robustness to illumination effects. Nonetheless, neither of these common feature representations is able to outperform the proposed SOE features with respect to the challenge of illumination changes.

In the *PETS* video, incorporating limited amounts of spatial organization information proves to be advantageous when the target is very small. In this case, the two SOE-based systems that perform some amount of spatial aggregation, **Frag-SOE** and **MS-SOE**, provide accurate tracking of the cyclist target until the end of the video. **PW-SOE**, **IVT**, **MIL**, and **VTD-Int** do not consider such spatial aggregation and all lose track of the target early in the sequence when the cyclist is partially occluded by a pedestrian. Interestingly, **MIL** appears to learn the appearance of the pedestrian and begins to track him or her as the sequence progresses (e.g., Frame 83).

Next, for the *Ming* video, **PW-SOE**, **IVT**, and **VTD-Int** provide comparably excellent target tracks throughout. In contrast, **MS-SOE** and **MIL** are able to follow the target to the completion of the video, but offer significantly inferior localization accuracy as compared to the top performers. Specifically, **MS-SOE** moves partially off target early in the sequence and allows the tracking window to grow much larger to incorporate more of the face and background than present in the initial template. Similarly, **MIL** drifts part-way off the individual's face at roughly Frame 225 and only provides coarse target localization thereafter. Additionally, **MIL** does not properly capture the scale changes as the distance from target to camera changes (e.g., Frame 925). Finally, the **Frag-SOE** struggles with this sequence in a similar manner to what transpired for *Sylvester*. Specifically, the target exhibits limited interframe motion and numerous changes in velocity. These weak and inconsistent velocity cues combined with a face target that is less distinct under local accumulation leads the tracker to seek matching regions on the untextured background.

For the challenging *Woman* sequence, of all the SOE-based and state-of-the-art systems considered in this qualitative analysis, only **PW-SOE** is capable of tracking the target throughout. The tracker is successful despite a moving camera in addition to extended partial occlusions with vehicles that share similar colors to the target itself. Although the track provided by **PW-SOE** is imperfect due to some unnecessary shearing of the crop box region, its performance is far superior to any of the alternatives, which all fall off the target by Frame 30 during the first partial occlusion.

In the *Car11* sequence, a moving camera is again encountered and this challenge causes **Frag-SOE** to fail early in the sequence and become attracted to other moving regions in the background with car taillight type appearances (e.g., Frame 200). The next tracker to experience difficulty is **MIL**, which begins to gradually drift off target starting around Frame 70 and eventually ends up in the oncoming lanes of traffic. **MS-SOE** and **VTD-Int** are the next trackers to drift off target at Frame 240 when the target begins to turn to the right around a corner. Only **PW-SOE** and **IVT** are capable of tracking the car target with comparable accuracy until the end of the sequence.

Finally, in *Pop Machines*, since the two individuals within the scene look very similar and walk closely to one another, **MIL** has difficulty distinguishing between them. For much of the video sequence, both **MIL** tracking windows are following the same individual. Additionally, **MS-SOE** succeeds at following one target to the end of the sequence, while **IVT** and

VTD-Int cannot surmount the full occlusions caused by the foreground pillar for either target. Only the feature representations used by **PW-SOE** and **Frag-SOE** that encompass both target dynamics and spatial organization allow for the differentiation between the two targets and success throughout this video.

In comparing the performance of the systems that employed the proposed SOE feature representation, the tradeoffs in maintaining spatial organization information (provided by **PW-SOE**, partially kept in **Frag-SOE**, and completely discarded in **MS-SOE**) are well documented. **Frag-SOE** and **MS-SOE** show a tendency to drift onto non-target locations that share similar feature characteristics with the target when aggregated over large support regions (e.g., occluding book and background computer monitor in *Occluded Face 2*, unstructured background in *Sylvester*, moving background regions in *Car11*). In contrast, **PW-SOE** does not exhibit these problems, as it maintains complete spatial organization of the features via its pointwise representation and the targets are distinguished from the non-target locations on that basis. On the other hand, the accumulation of feature information over a larger region did yield benefits when tracking very small targets, as demonstrated by the success of **Frag-SOE** and **MS-SOE** on the *PETS* sequence. In contrast, **PW-SOE** failed to track the target due the limited number of true target pixels within the tracking window, which is reduced further during the partial occlusion by the pedestrian. It appears that different tracking architectures may be more appropriate for specific scenarios, but the proposed SOE features provide a suitable base representation throughout.

3.3 Quantitative results

To evaluate the SOE-based systems versus other state-of-the-art algorithms and systems employing alternative feature representations, comparisons with ground truth data were made. Two standard metrics for single target visual tracking were used: the center location error (CLE) and the percentage of frames correctly tracked, given as VOC success rate (VOC-SR). The CLE computes the mean Euclidean distance between the ground truth center of mass and those provided by the various tracking systems across all video frames to yield an overall measure of tracking accuracy. VOC-SR is a measure based on the PASCAL VOC challenge [71]. For a given frame, the target is considered to be correctly tracked if its overlap with the ground truth bounding box is at least 50%, yielding an indication of the video percentage that tracking was "unsuccessful", due to poor localization or complete failure. These two complementary metrics were recently proposed for performance evaluation for the single target tracking problem [15] and have since been adopted as a standard for many recent evaluations [70], [36], [17].

Summary statistics using these two evaluation criterion are presented in Table 3 for all sixteen trackers listed in Table 2. Color-based trackers were restricted to the four sequences that had RGB data available. Since **MIL**, **A-BHMC**, and **VTD-Int**/**VTD-Color** are stochastic algorithms, they were executed five times and their errors averaged [12].

Absolute Performance. In considering Table 3, a number of meaningful trends can be observed. The top performers appear to be the trackers that leverage the spatiotemporal oriented energy representation, i.e., **PW-SOE**, **Frag-SOE**, **MS-SOE** as well as two of the state-of-the-art systems, **IVT** and **MIL**. Overall, the top performing tracker is the **PW-SOE** system, attaining best performance on three videos and second best on four videos under the CLE metric, as well as three best and one second best ratings under the VOC-SR measure. **IVT** is arguably the tracker that performs second best overall, achieving best performance on three videos under both evaluation criterion. Interestingly, whenever **PW-SOE** is outperformed by **IVT**, it was by a relatively small amount (i.e., 2.5 pixels for *Occluded Face 2*, 0.2 pixels for *Ming*, and 0.1 pixels for *Car11* under the CLE metric). In contrast, when **PW-SOE** excelled, **IVT** tended to experience more significant failures (e.g., *Sylvester*, *Illumination*, *Pop Machines*, and *Woman*). Further, in terms of the CLE and VOC-SR rankings, **Frag-SOE**, **MS-SOE** and **MIL** perform relatively on par, and are the best performers after **PW-SOE** and **IVT**. Thus, of the top five performers, three are SOE-based algorithms.

The additional state-of-the-art systems beyond **IVT** and **MIL** were less successful. The recent **A-BHMC** system did not offer a precise track for any sequence. **VTD-Int** and **VTD-Color** were able to significantly outperform **A-BHMC**, but the overall tracking accuracy of **VTD-Int** was otherwise not one of the best performers.

Given that the proposed SOE representation encompasses both appearance and dynamics, one might conjecture that for scenarios involving moving cameras, the SOE-based trackers will fail. Interestingly, the results in Table 3 for the two sequences with moving cameras, *Woman* and *Car11*, reveal that SOEs actually can perform very well with non-stationary cameras. In fact, for videos with smooth camera motion (panning, zooming, and looming) the **PW-SOE** system demonstrates that it can attain best and second best tracking accuracy amongst all 16 algorithms tested. However, it does appear that the inclusion of spatial layout information offered by the **PW** paradigm is critical for success when directly using SOE features that have been computed from the raw imagery. In particular, **Frag-SOE** and **MS-SOE** perform feature aggregation over larger spatial support regions, allowing them to be attracted to the clutter in the moving background more easily and eventually drift off target. Along these lines, an interesting direction for future research

Fig. 3. Qualitative tracking results for the test suite of videos. From top to bottom, left to right, the videos displayed are *Occluded Face 2*, *Tiger 2*, *Sylvester*, *Illumination*, *PETS*, *Ming*, *Woman*, *Car11*, and *Pop Machines*. Red, yellow, cyan, green, purple, and blue boxes correspond to tracking results for **PW-SOE**, **Frag-SOE**, **MS-SOE**, **IVT**, **MIL**, and **VTD-Int** respectively. For the *Pop Machines* video, two human targets are tracked where the results for the person starting on the right are shown with solid crop boxes and those for the person starting on the left are denoted via dashed boxes.

would be to explore the utility of image stabilization [72] as a preprocessing stage for **Frag-SOE** and **MS-SOE** to ameliorate the impact of camera motion.

Another interesting observation from the *Car11* video is that performance between **PW-SOE** and **PW-OE** are extremely comparable. The explanation is that when camera motion is present in a video, the target motion may become a less reliable component of the SOE representation; significantly, however, the **PW-SOE** tracker automatically still exploits the spatial appearance aspects of the SOEs to maintain good performance in such situations. Presumably, similar success was not achieved with the **PW-OE** tracker on the *Woman* sequence due to the highly non-rigid nature of the target. In this case, the SOE tracker is able to capitalize on the distinctive target motion information even in the presence of camera motion to achieve best results.

Performance vs. Feature Representation. An interesting trend observable from Table 3 with respect to features is that when holding the tracking paradigm constant (i.e., either **PW**, **Frag**, or **MS** architectures) and comparing the performance offered by the four different representations (i.e., SOE, OE, intensity, or color), SOEs generally outperform the alternatives. Specifically, SOEs demonstrated superior performance to the alternatives for 46 out of 62 comparisons for the CLE metric and 43 out of 62 comparisons for VOC-SR. This trend can also be observed in Table 3 by noting that the SOE variants of these tracking architectures achieve the most best and second best performance ratings on videos, as compared to their counterparts. Specifically, across all tracking paradigms and both metrics, SOE-based systems attained 22 bests or second bests, intensity-based systems yielded 4, while all other features had none.

Outside of the SOE-based systems, the only tracker that employs one of these paradigms and attains reasonable overall tracking accuracy is **Frag-Int** (second best performance on two videos under both metrics). This result is not surprising, as **Frag-Int** is a modern tracking architecture that was recently considered to be state-of-the-art [22] and is still used as a common baseline for comparison, e.g., [15], [70], [17].

Additional observations can be made regarding the feature representations when considering the **PW**, **Frag**, and **MS** paradigms in more detail. For instance, Table 3 reveals that illumination changes are problematic for the pure intensity features, as can be seen in the results for *Illumination*, which offer poor performance in the intensity-based variants of these algorithms. Bandpass filtering, (1), and normalization, (2), allow all SOE-based systems to track, relatively unaffected, through the pronounced illumination variations. Systems based on spatial oriented energies, although robust to illumination changes, are not sufficiently rich to track in this challenging sequence.

The *Tiger 2* video demonstrates that the purely spatially-based features (both raw intensity and spatial orientation) can be distracted easily by complicated cluttered scenery, especially when the target undergoes a slight change in appearance (e.g., partial occlusion by foliage, motion blur, opening of tiger's mouth). The addition of motion information in the SOE-based systems provides added discriminative power to avoid being trapped by clutter, leading to superior performance over the competing representations within the same tracking paradigm.

Also of interest is that when tracking extremely small targets in the *PETS* video, the richness of the SOE representation, capturing both appearance and dynamics, is required for success. The alternative representations that consider only appearance are not sufficient to follow such a small target, resulting in poor performance within every tracking paradigm.

Additionally, in *Pop Machines* the SOE representation is able to achieve success where the alternatives at least partially lose track of both targets. In this case, motion information is critical in distinguishing the targets, given their similar appearance. Relative success is had with the proposed SOE features within all tracking architectures, even as the targets cross paths and with the pillar providing further occlusion.

As the SOE features are computed from intensity images, it is important to include comparisons against color-based algorithms to ensure that omitting this potential source of information does not lead to significant performance degradation. The results in Table 3 indicate that inclusion of color information is not required to achieve high quality tracker output, nor does it guarantee that the resulting tracker will be robust. Indeed, in the entire evaluation, none of the color-based trackers achieved best or second best performance for any of the videos. Along these lines, it is of special interest to compare the performance of the **Frag**, **MS** and **VTD** trackers when the feature representation was changed from color to intensity. Here, the results show that overall tracking performance remains fairly comparable when switching between color and intensity within a paradigm. **Frag-Color** sees the most advantage in employing color, but still does not become competitive with the top performers.

Overall, comparisons of the various feature representations show that tracking performance is very sensitive to which features are employed, even when other components of the tracker are held constant. The specific choice of feature representation is critical in overcoming certain challenges in tracking including, illumination changes, clutter, occlusion, appearance changes, and multiple targets with similar appearance. Generally, the proposed spatiotemporal oriented energy features offer superior tracking accuracy over extant features with respect to these challenges.

In summary, the empirical evaluation has demonstrated three main findings regarding the 16 trackers that were compared. (i) One of the proposed trackers, **PW-SOE**, provides overall best tracking accuracy and outperforms the state-of-the-art systems on the test suite. The two additional SOE-based systems **Frag-SOE** and **MS-SOE**, demonstrated overall performance within the top five trackers and outperformed several strong baselines. These results indicate that SOE features can be used for tracking to attain state-of-the-art accuracy. (ii) SOE features have been experimentally validated as being amenable and effective in three distinct region tracking paradigms that maintain varying amount of spatial layout information, illustrating the representation's flexibility and breadth of applicability. (iii) The richness combined with the robustness to illumination allowed the proposed SOE representation to outperform the alternative intensity, color, and spatial orientation features. In a direct comparison, the SOE features outperformed some of the most commonly-used features in tracking today with respect to several of the outstanding challenges in visual tracking.

4 Discussion and Summary

The main contribution of this paper is the novel application of a spatiotemporal oriented energy representation to the field of visual tracking. The energy-based representation is shown to be effective and broadly applicable in a range of representative region trackers that incorporate differing amounts of target spatial layout information during tracking. The proposed SOE representation uniformly captures both the spatial and temporal characteristics of a target while remaining robust to illumination. The feature set is uncommonly rich, supporting tracking through appearance and illumination changes, erratic target motion, moving cameras, complicated backgrounds, and occlusions. Indeed, in empirical comparisons to alternative commonly employed features and extant state-of-the-art trackers, the proposed approach has been shown to yield exceptionally strong performance in response to such challenges.

Acknowledgments

This research was supported by an NSERC Discovery grant to R. Wildes.

TABLE 3
Summary of quantitative results for all 16 trackers. Each entry in the table is in the format CLE (pixels) / VOC-SR (%). Green and red indicate best and second best performance, respectively.

Algorithm	*Occl. Face 2*	*Tiger 2*	*Sylv.*	*Illum.*	*PETS*	*Ming*	*Woman*	*Car11*	*Pop Mach.*
PW-SOE	**8.5**/90.2	21.8/21.1	**8.2**/**88.6**	**8.2**/**96.7**	35.0/15.6	**3.3**/98.3	**11.1**/**73.2**	**2.6**/**90.8**	**16.0**/40.1
PW-OE	18.2/57.7	32.5/15.5	81.2/36.5	116.0/10.0	36.3/8.9	3.9/96.6	113.6/21.1	3.3/84.2	86.7/6.2
PW-Int	27.2/50.3	46.2/15.5	50.6/35.7	91.1/33.3	48.8/4.4	12.7/59.7	93.7/16.9	35.8/48.7	41.7/24.3
Frag-SOE	54.1/27.6	**17.5**/**43.7**	82.3/0.8	**7.3**/**93.3**	**2.7**/**62.2**	127.4/0.7	131.5/2.4	88.6/2.6	**16.9**/41.9
Frag-OE	37.1/47.2	31.2/19.7	65.9/22.8	122.1/13.3	35.2/22.2	99.5/13.9	126.0/6.0	31.0/69.7	48.3/38.5
Frag-Int	20.0/68.1	74.8/18.3	**9.2**/**81.0**	113.2/33.3	44.9/26.7	10.5/69.2	**80.8**/**25.9**	44.8/15.8	105.4/18.5
Frag-Color	—	—	—	12.2/83.3	20.8/17.8	—	128.1/0.4	22.0/31.6	—
MS-SOE	77.2/1.2	30.9/16.9	18.4/22.1	8.5/**96.7**	**3.2**/**77.8**	34.3/0.3	133.8/1.6	20.8/55.3	29.28/**45.2**
MS-OE	78.2/1.8	77.2/1.4	23.4/40.3	136.4/3.3	49.1/4.4	13.9/55.3	94.0/35.0	26.2/13.2	114.3/6.2
MS-Int	29.4/42.3	47.6/4.2	24.5/27.0	51.9/36.7	49.2/4.4	20.3/32.5	140.3/3.5	28.6/39.5	107.2/11.3
MS-Color	—	—	—	124.4/10.0	50.0/6.7	—	136.2/0.2	26.9/23.7	—
IVT	**6.3**/**98.2**	38.7/21.1	91.7/44.1	115.3/16.7	43.0/20.0	**3.1**/**99.7**	269.2/3.8	**2.5**/**100.0**	55.8/32.7
MIL	19.0/86.7	**11.2**/**61.4**	13.1/68.4	28.2/62.7	27.7/18.7	18.7/36.3	137.1/4.4	40.9/18.7	38.6/**60.8**
A-BHMC	—	—	—	95.5/8.7	40.6/1.3	—	82.4/0.5	38.9/5.3	—
VTD-Int	9.8/**97.7**	22.5/40.0	15.9/79.6	28.8/28.7	31.6/24.0	4.1/**98.7**	132.6/4.6	22.7/60.3	55.2/32.4
VTD-Color	—	—	—	34.5/26.7	31.8/26.2	—	135.6/4.3	28.3/56.6	—

REFERENCES

[1] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Computing Surveys*, vol. 38, no. 4, pp. 1–45, 2006.

[2] K. Cannons, "A review of visual tracking," Technical Report CSE-2008-07, York University, Department of Computer Science and Engineering, 2008.

[3] S. S. Intille, J. W. Davis, and A. F. Bobick, "Real-time closed-world tracking," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1997.

[4] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: Real-time tracking of the human body," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 780–785, 1997.

[5] C. Stauffer and W. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 747–757, 2000.

[6] Z. Yin and R. Collins, "Belief propagation in a 3D spatio-temporal MRF for moving object detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2007.

[7] V. Reilly, H. Idrees, and M. Shah, "Detection and tracking of large number of targets in wide area surveillance," in *Proc. 11th European Conf. Computer Vision*, 2010, pp. 186–199.

[8] B. Song, T. Y. Jeng, E. Staudt, and A. K. Roy-Chowdhury, "A stochastic graph evolution framework for robust multi-target tracking," in *Proc. 11th European Conf. Computer Vision*, 2010, pp. 605–619.

[9] C. Huang, B. Wu, and R. Nevatia, "Robust object tracking by hierarchical association of detection responses," in *Proc. 10th European Conf. Computer Vision*, 2008, pp. 788–801.

[10] A. Andriyenko and K. Schindler, "Globally optimal multi-target tracking on a hexagonal lattice," in *Proc. 11th European Conf. Computer Vision*, 2010, pp. 466–479.

[11] H. Grabner, C. Leistner, and H. Bischof, "Semi-supervised on-line boosting for robust tracking," in *Proc. 10th European Conf. Computer Vision*, 2008, pp. 234–247.

[12] B. Babenko, M. Yang, and S. Belongie, "Visual tracking with online multiple instance learning," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009.

[13] Z. Kalal, J. Matas, and K. Mikolajczyk, "P-N learning: Bootstrapping binary classifiers by structural constraints," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010.

[14] A. Saffari, M. Godec, T. Pock, and C. Leistner, "Online multi-class LPBoost," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010.

[15] J. Santner, C. Leistner, A. Saffari, T. Pock, and H. Bischof, "PROST: Parallel robust online simple tracking," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2010, pp. 723–730.

[16] B. Zeisl, C. Leistner, A. Saffari, and H. Bischof, "On-line semi-supervised multiple-instance boosting," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010.

[17] R. Yao, Q. Shi, C. Shen, Y. Zhang, and A. Hengel, "Robust tracking with weighted online structured learning," in *Proc. 12th European Conf. Computer Vision*, 2012.

[18] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2005.

[19] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2000.

[20] G. Hager, M. Dewan, and C. Stewart, "Multiple kernel tracking with SSD," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2004.

[21] S. Birchfield and S. Rangarajan, "Spatiograms versus histograms for region-based tracking," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2005.

[22] A. Adam, E. Rivlin, and I. Shimshoni, "Robust fragments-based tracking using the integral histogram," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2006.

[23] K. Cannons and R. Wildes, "Spatiotemporal oriented energy features for visual tracking," in *Proc. Asian Conf. Computer Vision*, 2007, pp. 532–543.

[24] F. Wang, S. Yu, and J. Yang, "A novel fragments-based tracking algorithm using mean shift," in *Proc. Int'l Conf. Control, Automation, Robotics and Vision*, 2008.

[25] K. Okuma, A. Taleghani, N. D. Freitas, J. Little, and D. Lowe, "A boosted particle filter: Multitarget detection and tracking," in *Proc. 8th European Conf. Computer Vision*, 2004, pp. 28–39.

[26] B. Lucas and T. Kanade, "An iterative image registration technique with application to stereo vision," in *Proc. DARPA Image Understanding Workshop*, 1981, pp. 121–130.

[27] F. G. Meyer and P. Bouthemy, "Region-based tracking using affine motion models in long image sequences," *CVGIP: Image Understanding*, vol. 60, no. 2, pp. 119–140, 1994.

[28] D. Beymer and K. Konolige, "Real-time tracking of multiple people using continuous detection," in *Proc. 7th IEEE Int'l Conf. Computer Vision*, 1999.

[29] H. T. Nguyen, M. Worring, and R. Boomgaard, "Occlusion robust adaptive template tracking," in *Proc. 8th IEEE Int'l Conf. Computer Vision*, 2001, pp. 678–683.

[30] A. Jepson, D. Fleet, and T. El-Maraghi, "Robust online appearance models for visual tracking," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, no. 10, pp. 1296–1311, 2003.

[31] D. A. Ross, J. Lim, and R. S. Lin, "Incremental learning for robust visual tracking," *Int'l J. Computer Vision*, vol. 77, pp. 125–141, 2008.

[32] J. Kwon and K. M. Lee, "Tracking of a non-rigid object via patch-based dynamic appearance modeling and adaptive

basin hopping Monte Carlo sampling," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009.

[33] ——, "Visual tracking decomposition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010.

[34] J. R. Bergen, P. Anandan, K. J. Hanna, and R. Hingorani, "Hierarchical model-based motion estimation," in *Proc. Second European Conf. Computer Vision*, 1992, pp. 237–252.

[35] M. J. Black and P. Anandan, "The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields," *Computer Vision and Image Understanding*, vol. 63, no. 1, pp. 75–104, 1996.

[36] S. Oron, A. Bar-Hillel, D. Levi, and S. Avidan, "Locally orderless tracking," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2012.

[37] P. Anandan, "A computational framework and an algorithm for the measurement of visual motion," *Int'l J. Computer Vision*, vol. 2, no. 3, pp. 283–310, 1989.

[38] P. Fieguth and D. Terzopoulos, "Color-based tracking of heads and other mobile objects at video frame rates," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1997.

[39] S. M. S. Nejhum, J. Ho, and M. H. Yang, "Visual tracking with histograms and articulating blocks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2008.

[40] I. Haritaoglu and M. Flickner, "Detection and tracking of shopping groups in stores," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2001.

[41] Y. Bogomolov, G. Dror, S. Lapchev, E. Rivlin, and M. Rudzsky, "Classification of moving targets based on motion and appearance," in *Proc. British Machine Vision Conf.*, 2003, pp. 142–149.

[42] D. Cremers and C. Schnörr, "Statistical shape knowledge in variational motion segmentation," *Image and Vision Computing*, vol. 21, no. 1, pp. 77–86, 2003.

[43] K. Sato and J. Aggarwal, "Temporal spatio-velocity transformation and its application to tracking and interaction," *Computer Vision and Image Understanding*, vol. 96, no. 2, pp. 100–128, 2004.

[44] S. Birchfield, "Elliptic head tracking with intensity gradients and color histograms," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1998.

[45] A. Elgammal, R. Duraiswami, and L. Davis, "Probabilistic tracking in joint feature-spatial spaces," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2003.

[46] C. Yang, R. Duraiswami, and L. Davis, "Efficient mean-shift tracking via a new similarity measure," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2005.

[47] S. Avidan, "Ensemble tracking," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 2, pp. 261–271, 2007.

[48] Z. Yin, F. Porikli, and R. T. Collins, "Likelihood map fusion for visual object tracking," in *Proc. IEEE Workshop Applications of Computer Vision*, 2008.

[49] T. Brox, B. Rosenhahn, J. Gall, and D. Cremers, "Combined region and motion-based 3D tracking of rigid and articulated objects," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, pp. 402–415, 2010.

[50] D. Heeger, "Optical flow from spatiotemporal filters," *Int'l J. Computer Vision*, vol. 1, no. 4, pp. 297–302, 1988.

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

[52] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. IEEE Int'l Workshop Performance Evaluation of Tracking and Surveillance*, 2005, pp. 65–72.

[53] K. Derpanis, M. Sizintsev, K. Cannons, and R. Wildes, "Efficient action spotting based on a spacetime oriented structure representation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010.

[54] K. Derpanis and R. Wildes, "Dynamic texture recognition based on distributions of spacetime oriented structure," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010.

[55] R. Wildes and J. Bergen, "Qualitative spatiotemporal analysis using an oriented energy representation," in *Proc. 6th European Conf. Computer Vision*, 2000, pp. 784–796.

[56] M. Sizintsev and R. Wildes, "Spatiotemporal stereo via spatiotemporal quadric element (stequel) matching," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009.

[57] K. J. Cannons, J. M. Gryn, and R. P. Wildes, "Visual tracking using a pixelwise spatiotemporal oriented energy representation," in *Proc. 11th European Conf. Computer Vision*, 2010, pp. 511–524.

[58] E. Adelson and J. Bergen, "Spatiotemporal energy models for the perception of motion," *J. Optical Soc. Am. A*, vol. 2, no. 2, pp. 284–299, 1985.

[59] T. Lindeberg, *Scale-Space Theory in Computer Vision*. Kluwer Academic Publishers, 1994.

[60] M. Brand and V. Kettnaker, "Discovery and segmentation of activities in video," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 844–851, 2000.

[61] W. Freeman and E. Adelson, "The design and use of steerable filters," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 13, no. 9, pp. 891–906, 1991.

[62] A. Zaharescu and R. Wildes, "Anomalous behaviour detection using spatiotemporal oriented energies, subset inclusion histogram comparison and event-driven processing," in *Proc. 11th European Conf. Computer Vision*, 2010.

[63] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distribution," *Bulletin of the Calcutta Mathematical Society*, vol. 35, pp. 99–110, 1943.

[64] J. Shi and C. Tomasi, "Good features to track," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1994.

[65] P. Huber, *Robust Statistical Procedures*. SIAM Press, 1977.

[66] R. Collins, "Mean-shift blob tracking through scale space," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2003.

[67] Z. Zivkovic and B. Krose, "An EM-like algorithm for color-histogram tracking," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2004.

[68] PETS, http://www.cvg.rdg.ac.uk/PETS2001/, 2001.

[69] D. Ross, J. Lim, and M. Yang, "Adaptive probabilistic visual tracking with incremental subspace update," in *Proc. 8th European Conf. Computer Vision*, 2004.

[70] X. Jia, H. Lu, and M. Yang, "Visual tracking via adaptive structural local sparse appearance model," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2012.

[71] M. Everingham, L. V. Gool, C. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int'l J. Computer Vision*, vol. 88, pp. 303–338, 2010.

[72] P. J. Burt, J. R. Bergen, R. Hingorani, W. A. Lee, and A. Leung, "Object tracking with a moving camera," in *Proc. IEEE Workshop Visual Motion*, 1989, pp. 2–12.

Kevin Cannons received the BSc (Honours with Distinction) degree in computer engineering from the University of Manitoba, Canada, in 2003, the MASc degree in computer engineering from the University of Toronto, Canada, in 2005 and the PhD degree in computer science from York University, Canada, 2011. Currently, he is a postdoctoral researcher in the School of Computing Science at Simon Fraser University. His major field of interest is computer vision with specific emphasis on visual tracking and spatiotemporal analysis.

Richard Wildes (Member, IEEE) received the PhD degree from the Massachusetts Institute of Technology in 1989. Subsequently, he joined Sarnoff Corporation in Princeton, New Jersey, as a Member of the Technical Staff in the Vision Technologies Group. In 2001, he joined the Department of Computer Science and Engineering at York University, Toronto, where he is an Associate Professor and a member of the Centre for Vision Research. Honours include receiving a Sarnoff Corporation Technical Achievement Award, the IEEE D.G. Fink Prize Paper Award for his Proceedings of the IEEE publication Iris recognition: An emerging biometric technology and twice giving invited presentations to the US National Academy of Sciences. His main areas of research interest are computational vision, as well as allied aspects of image processing, robotics and artificial intelligence.