Visual Tracking Using a Pixelwise Spatiotemporal Oriented Energy Representation

Kevin J. Cannons, Jacob M. Gryn, and Richard P. Wildes

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

Abstract. This paper presents a novel pixelwise representation for visual tracking that models both the spatial structure and dynamics of a target in a unified fashion. The representation is derived from spatiotemporal energy measurements that capture underlying local spacetime orientation structure at multiple scales. For interframe motion estimation, the feature representation is instantiated within a pixelwise template warping framework; thus, the spatial arrangement of the pixelwise energy measurements remains intact. The proposed target representation is extremely rich, including appearance and motion information as well as information about how these descriptors are spatially arranged. Qualitative and quantitative empirical evaluation on challenging sequences demonstrates that the resulting tracker outperforms several alternative state-of-the-art systems.

1 Introduction

Tracking of objects in image sequences is a well-studied problem in computer vision that has seen numerous advances over the past thirty years. There are several direct applications of "following a target" (e.g., surveillance and active camera systems); furthermore, many computer vision problems rely on visual trackers as an initial stage of processing (e.g., activity and object recognition). Between the direct applications of target tracking and the evolution of visual tracking into a basic stage for subsequent processing, there is no shortage of motivation for the development of robust visual trackers.

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.

K. Daniilidis, P. Maragos, N. Paragios (Eds.): ECCV 2010, Part IV, LNCS 6314, pp. 511–524, 2010.

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 because as one component of the representation experiences a fast change, other components may remain more consistent. In the current approach, a pixelwise spatiotemporal oriented energy representation is employed. 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.

Visual trackers can be coarsely divided into three general categories: (i) discrete feature trackers (ii) contour-based trackers, and (iii) region-based trackers [9]. 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 subsequent data association between the detected foreground "blobs" [30,27,28]. Another subclass of region trackers collapses the spatial information across the target support and uses a histogram representation of the target during tracking [11,16,5,10]. The work in [10] presents the most relevant histogram tracker to the current approach because both share a similar energy-based feature set. A final subcategory of region-based trackers retains spatial organization within the tracked area by using (dense) pixelwise feature measurements. Various feature measurements have been considered [25,13,20,23]. Further, several approaches have been developed for updating/adapting the target representation on-line [19,20,23,3]. This final subcategory of region trackers is of relevance to the present work, as it maintains a representation of dense pixelwise feature measurements with a parameterized model of target motion.

Throughout all categories of trackers, a relatively under-researched topic is that of identifying an effective representation that models both the spatial and dynamic properties of a target in a uniform fashion. While some tracking-related research has sought to combine spatial and motion-based features (e.g., [7,24]), the two different classes of features are derived separately from the image sequence, which has potential for making subsequent integration challenging. A single exception is the tracker noted above that derived its features from spatiotemporal oriented energy measurements, albeit ultimately collapsing across spatial support [10], making it more susceptible to clutter (e.g., background and foreground share similar overall feature statistics, yet would be distinguished by spatial layout) and "blind" to more complex motions (e.g., rotation).

In light of previous research, the main contributions of the present paper are as follows. (i) A novel oriented energy representation that retains the spatial organization of the target is developed for visual tracking. Although similar oriented energy features have been used before in visual trackers [10] and other areas of image sequence processing (e.g., [2,15,29,26,12,31]), it appears that these features have never been deployed in a pixelwise fashion to form the fundamental features for tracking. (ii) A method is derived for instantiating this representation within a parametric flow estimation tracking algorithm. (iii) The discriminative power of the pixelwise oriented energy representation is demonstrated via a direct comparison against other commonly-used features. (iv) The overall tracking implementation is demonstrated to perform better than several state-of-the-art algorithms on a set of challenging video sequences during extensive qualitative and quantitative comparisons.

2 Technical Approach

2.1 Features: Spatiotemporal Oriented Energies

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 [2]. These energies are derived from the filter responses of orientation selective bandpass filters that are applied to the spatiotemporal volume representation of a video stream. 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., image 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 aforementioned energies are well-suited to form the feature representation in visual tracking applications for four significant reasons. (i) A rich description of the target is attained due to the fact that oriented energies encompass both target appearance and dynamics. This richness allows for a tracker that is more robust to clutter both in the form of background static structures and other moving targets in the scene. (ii) The oriented energies are robust to illumination changes. By construction, the proposed feature set provides invariance to both additive and multiplicative image intensity changes. (iii) 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). In a complementary fashion, coarser scales provide information regarding the overall target velocity and its gross shape. (iv) The representation is efficiently implemented via linear and pixelwise non-linear operations [14], with amenability to real-time realizations on GPUs [31].

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 within a Gaussian pyramid [14]. To attain an initial measure of energy, the filter responses are pixelwise rectified (squared) and summed according to

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

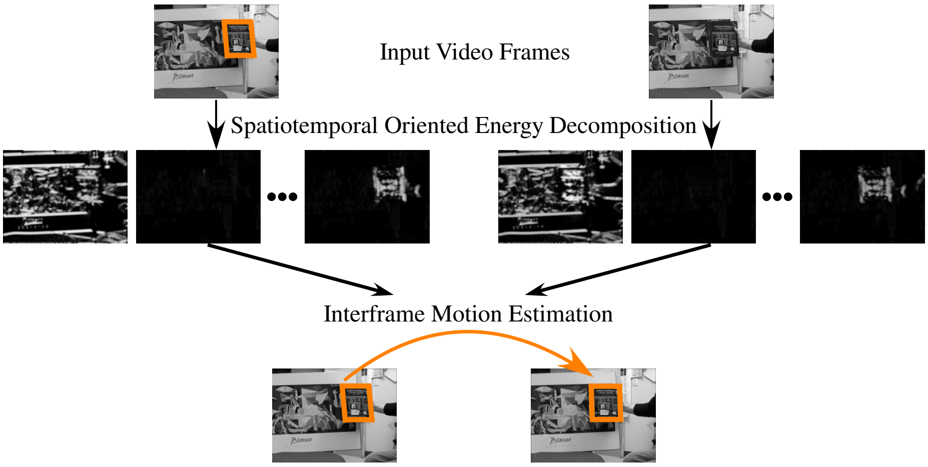
Input Video Frames
Spatiotemporal Oriented Energy Decomposition
Interframe Motion Estimation

Fig. 1. Overview of visual tracking approach. (top) Two frames from a video where a book is being tracked. The left image is the first frame with a crop box defining the target's initial location. The right image is a subsequent frame where the target must be localized. (middle) Spacetime oriented filters decompose the input into a series of channels capturing spatiotemporal orientation; left-to-right the channels for each frame correspond roughly to horizontal static structure, rightward and leftward motion. (bottom) Interframe motion is computed using the oriented energy decomposition.

where $\mathbf{x} = (x, y, t)$ are spatiotemporal image coordinates, I is an image, and $*$ denotes the convolution operator. It is the bandpass nature of the G_2 and H_2 filters during the computation of (1) that leads to the energies' invariance to additive image intensity variations.

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, pixelwise normalization is performed,

$$\hat{E}\left(\mathbf{x};\theta,\gamma\right) = \frac{E\left(\mathbf{x};\theta,\gamma\right)}{\sum_{\tilde{\gamma}}\sum_{\tilde{\theta}} E\left(\mathbf{x};\tilde{\theta},\tilde{\gamma}\right) + \epsilon} \ , \tag{2}$$

where ϵ is a constant introduced as a noise floor and to avoid numerical instabilities when the overall energy content is small. Additionally, the summations, (2), consider all scales and orientations at which filtering is performed (here, the convention is to use $\tilde{}$ for variables of summation). The representation's invariance to multiplicative intensity changes is a direct result of this normalization, (2).

2.2 Target Representation

Depending on the tracking architecture being employed, pixelwise energy measurements, (2), can be manipulated to define various target representations (e.g., collapsed to form an energy histogram, parameterized by orientation and scale [10]). The present approach retains the target's spatial organization by defining the representation in terms of a pixelwise template for tracking based on parametric registration of the template to the image across a sequence [21,4,6]. In particular, the template is initially defined as

$$T(x, y, \theta, \gamma) = \hat{E}(x, y, t_0; \theta, \gamma) \tag{3}$$

for energies measured at some start time, t_0, and spatial support, (x, y), over some suitably specified region. 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. It will be shown in Sec. 3 that retaining pixelwise organization leads to significantly better performance than collapsing over target support, as in [10].

As an illustrative example, Fig. 1 shows a sample oriented energy decomposition where a book is moving to the left in front of a cluttered background. Consideration of the first channel shows that it responds strongly to horizontal static structures both on the book and in the background. The second channel corresponds roughly to rightward motion and as such, results in negligible energy across the entire image frame. Significantly, note how the third channel tuned roughly to leftward motion yields strong energy responses on the book and small responses elsewhere, effectively differentiating between target and background.

Finally, note that the target representation, (3), is in contrast to standard template tracking-based systems that typically only utilize a single channel of intensity features during estimation [4,6]. Further, even previous approaches that have considered multiple measurements/pixel make use of only spatially derived features (e.g., [20]), which will be shown in Sec. 3 to significantly limit performance in comparison to the current approach.

2.3 Robust Motion Estimation

Tracking using a pixelwise template approach consists of matching the template, T, 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, T, and the image frame, I, are represented in terms of oriented energy measurements, (2). To illustrate the efficacy of this representation, 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 [25,22,23]. Further, the optical flow constraint equation (OFCE) [17] is used to formulate a match measure between features (oriented energies) that are aligned by the motion model. Under a parametric model, the OFCE can be written as

$$\nabla^\top \hat{E} \mathbf{u}(\mathbf{a}) + \hat{E}_t = 0 \ , \tag{4}$$

where $\nabla^\top \hat{E} = \left(\hat{E}_x, \hat{E}_y\right)$ are the first-order spatial derivatives of the image energy measurements, (2), for some specific orientation, θ, and scale, γ, $\hat{E}_t$ is the first order temporal derivative, $\mathbf{u} = (u, v)^{\top}$ is the flow vector, and $\mathbf{a} = (a_0, a_1, \ldots a_5)^{\top}$ are the six affine motion parameters for the local region. The affine motion model is explicitly defined as

$$\mathbf{u}\,(x, y; \mathbf{a}) = (a_0 + a_1 x + a_2 y, a_3 + a_4 x + a_5 y)^{\top} \,. \tag{5}$$

The affine parameters, $\mathbf{a}$, are estimated by minimizing the error in the constraint equation, (4), summed over the target support. Significantly, in the present approach the target representation spans not just a single image plane, but multiple feature channels (orientations and scales) of spatiotemporal oriented energies. As a result, the error minimization is performed across the target support and over all feature channels. To measure deviation from the optical flow constraint, a robust error metric, $\rho(\eta, \sigma)$, is utilized [6]. The robust metric is beneficial for occlusion events, imprecise target delineations that include background pixels, and target motion that deviates from the affine motion model (e.g., non-rigid, articulated motion). With the above considerations in mind, the affine motion parameters, defining the interframe target motion, are taken as

$$\arg\min_{\mathbf{a}} \sum_{\tilde{\mathbf{x}}} \sum_{\tilde{\theta}} \sum_{\tilde{\gamma}} \rho \left[\nabla^{\top} \hat{E} \left(\tilde{x}, \tilde{y}, t; \tilde{\theta}, \tilde{\gamma} \right) \mathbf{u}\,(\mathbf{a}) + \hat{E}_t \left(\tilde{x}, \tilde{y}, t; \tilde{\theta}, \tilde{\gamma} \right), \sigma \right] , \tag{6}$$

where summations are across target support, $\tilde{\mathbf{x}} = (\tilde{x}, \tilde{y})$, as well as all feature channel orientations, $\tilde{\theta}$, and scales, $\tilde{\gamma}$. In the present implementation, the Geman-McClure error metric [18] is utilized with σ the robust metric width, as suggested in [6]. The minimization to yield the motion estimate, (6), is performed using a gradient descent procedure [6]. To increase the capture range of the tracker, the minimization process is performed in a coarse-to-fine fashion [4,6].

The affine parameters estimated using (6) bring the target template into alignment with the closest matching local set of oriented energy features that are derived from the current image frame. At the conclusion of processing each frame, the position of the target is updated via the affine motion estimates, $\mathbf{a}$, forming a track of the target across the video sequence. After target positional updating has been completed, the next video frame is obtained and the motion estimation process between the template and the new image data is performed. This process is repeated until the end of the video is reached.

2.4 Template Adaptation

Tracking algorithms require a means of ensuring that the internal target representation (i.e., the template) remains up-to-date with the true target characteristics in the current frame, especially when tracking over long sequences. For the proposed tracker, 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. In the present implementation, a simple template update scheme is utilized that computes a weighted combination of the aligned, optimal candidate oriented energy image features in the current frame, C^i, and the previous template

$$T^{i+1}(\mathbf{x}, \theta, \gamma) = \alpha T^i(\mathbf{x}, \theta, \gamma) + (1 - \alpha) C^i(\mathbf{x}, \theta, \gamma) \quad , \tag{7}$$

where α is a constant adaptation parameter controlling the rate of the updates (c.f., [19]). Although this update mechanism is far from the state-of-the-art in adaptation [20,23,3], the implementation achieves competitive results due to the overall strength of the pixelwise oriented energy feature set.

To summarize, Fig. 1 provides an overview of the entire system.

3 Empirical Evaluation

Three experiments were performed on the resulting system to assess its ability to track affine deformations, determine the power of the pixelwise spatiotemporal oriented energy representation, and compare its performance against alternative trackers. For all three experiments, unless otherwise stated, the following parameters were used. For the representation, 10 orientations were selected 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. Energies were computed at a single scale, corresponding to direct application of the oriented filters to the input imagery. For motion estimation, coarse-to-fine processing operated over 4 levels of a Gaussian pyramid built on top of the oriented energy measurements. Templates were hand initialized and updated with $\alpha \approx 0.9$. Video results for all experiments are available in supplemental material and online [1].

Experiment 1: Tracking affine deformations. This experiment illustrates the ability of the proposed system to estimate a wide range of affine motions when tracking a planar target (book) against a similarly complicated texture background; see Fig. 2. The target undergoes severe deformations including significant rotation, shearing, and scaling. While other feature representations (e.g., pixel intensities) also might perform well in these cases, the experiment documents that the spatiotemporal oriented energy approach, in particular, succeeds when experiencing affine deformations and that the motion estimator itself is capable of achieving excellent performance.

It is seen that the system accurately tracks the book throughout all tested cases. Performance decreases slightly when the book undergoes significant rotation. This drop is not alarming because part of the oriented energy representation encompasses the spatial orientation of the target, which is clearly changing during a rotation. Despite this "appearance change" (under the proposed feature representation), success is had for three reasons. (i) Template updates adjust the internal template representation to more accurately represent the target in the current frame. (ii) The filters used in computing the oriented energies (G_2 and H_2) are broadly tuned and allow for inexact matches. (iii) The tracker can utilize other aspects of the rich representation (e.g., motion) that remain

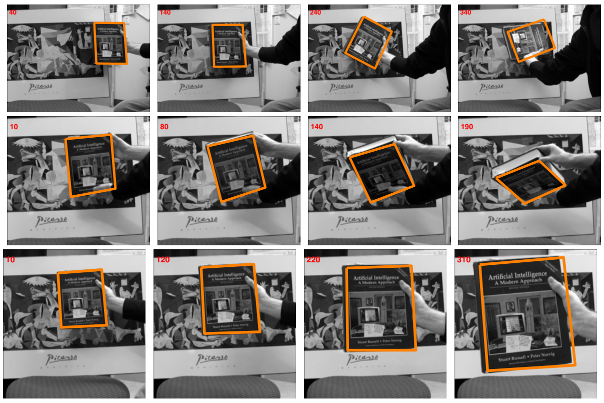

Fig. 2. Tracking through affine deformations. (row 1) Translation and subsequent rotation as the book rotates in plane over 90°. (row 2) Shearing as book rotates significantly out-of-plane. (row 3) Scaling as book moves toward camera. Orange box indicates tracked target.

relatively constant throughout the rotation. It is expected that if a more elaborate template update scheme were used, the results could be further improved.

Experiment 2: Feature set comparison. This experiment provides a comparison between the proposed spatiotemporal oriented energy features and two alternatives. In all cases, the motion estimation and template updates are identical (i.e., according to Sections 2.3 and 2.4). The single differentiating factor is the feature set. In the first system, pixelwise spatiotemporal oriented energies were used (10 orientations, as above) while the second tracker simply employed pixelwise raw image intensities. The third feature representation was purely spatial oriented energies, computed at four orientations (0°, 45°, 90°, and 135°), so as to span the space of 2D orientations for the highest order filters.

Five difficult, publicly available video sequences were used to demonstrate the points of this experiment. All five videos with ground truth can be downloaded [1,3]. The videos are documented in Table 1; results are shown in Fig. 3. To

Table 1. Experiments 2 and 3 video documentation. See Figs. 3 and 4 for images.

Occluded Face 2 [3]: Facial target. In plane target rotation. Cluttered background. Appearance change via addition of hat. Significant occlusion by book and hat.
Sylvester [23]: Hand-held stuffed animal target. Fast erratic motion, including out-of-plane rotation/shear. Illumination change across trajectory.
Tiger 2 [3]: Hand-held stuffed animal target. Small target with fast erratic motion. Cluttered background/foreground. Occlusion as target moves amongst leaves.
Ming [23]: Facial target. Variable facial expression. Fast motion. Significant illumination change across trajectory.
Pop Machines [original]: Similar appearing targets with crossing trajectories. Low quality surveillance video. Harsh lighting. Full occlusion from central pillar.

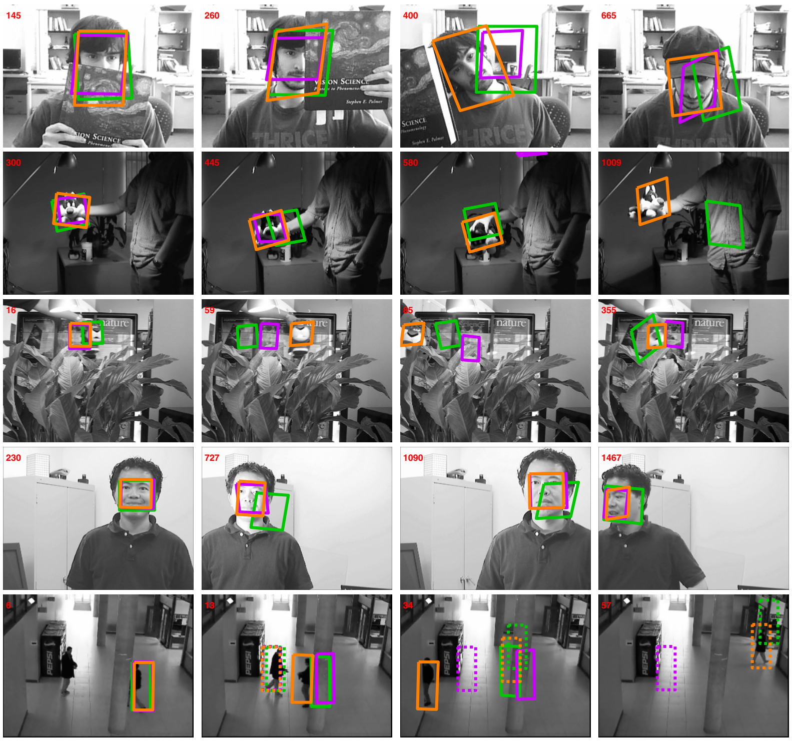

Fig. 3. Feature comparisons. Frame numbers are shown in the top left corner of each image. Top-to-bottom by row, shown are *Occluded Face 2*, *Sylvester*, *Tiger 2*, *Ming*, and *Pop Machines* of Table 1. Orange, purple, and green boxes are for spatiotemporal oriented, purely spatial oriented, and raw intensity features, resp.

ensure fair comparison with previous literature, the initial tracking boxes were set to the ground truth coordinates for the selected start frames, where available. For *Pop Machines*, trackers were initialized at the onset of each target's motion.

This second experiment clearly illustrates the fact that the choice of feature representation is critical in overcoming certain challenges in tracking including, illumination changes, clutter, appearance changes, and multiple targets with similar appearance. With reference to Fig. 3, illumination changes are problematic for the pure intensity features, as can be seen in the results from the *Ming* (Frames 727 and 1090) and *Sylvester* (Frame 445) sequences. Bandpass filtering, (1), and normalization, (2), allows both energy-based approaches to track, relatively unaffected, through these illumination variations. The *Occluded Face 2* (Frame 400) and *Tiger 2* (Frames 59 and 85) videos demonstrate that the purely spatially-based features (both raw intensity and orientation) can easily be distracted by complicated cluttered scenery, especially when the target undergoes a slight change in appearance (e.g., rotation of head, partial occlusion by foliage, motion blur). The addition of motion information in the spatiotemporal approach provides added discriminative power to avoid being trapped by clutter. Appearance changes caused by out of plane rotations in *Sylvester* (Frame 580) and the addition of a hat in *Occluded Face 2* (Frame 665) are also problematic for the raw intensity features and the spatial oriented energies; however, motion information allows the spatiotemporal approach to succeed during the appearance changes. Notice also that when the motion changes rapidly (*Sylvester*, *Tiger 2*), the spatiotemporal approach can still maintain track, as the spatial component of the representation remains stable while the motion component adapts via update, (7). Finally, in *Pop Machines* the spatiotemporal energy 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. Success is had with the proposed approach even as the targets cross paths and with the pillar providing further occlusion.

Experiment 3: Comparison against alternative trackers. In this experiment, analyses are conducted that show the proposed spatiotemporal oriented energy tracker (**SOE**) meets or exceeds the performance of several alternative strong trackers. The trackers considered are the multiple instance learning tracker (**MIL**) [3], the incremental visual tracker (**IVT**) [23], and a tracker that uses a similar oriented energy representation, but that is spatially collapsed across target support to fit within the mean shift framework (**MS**) [10]. The parameters for the competing algorithms were assigned values that were recommended by the authors or those that provided superior results. The videos used for this experiment are the same ones used in Exp. 2.

Figure 4 shows qualitative tracker results. For *Occluded Face 2*, **SOE** and **IVT** provide very similar qualitative results; whereas, **MIL** becomes poorly localized during the later stages of the video. The collapsing of spatial arrangement information in conjunction with a loose initial target window limits the performance of **MS**, as it is distracted onto the background. Also problematic for **MIL** and **MS** is that they only estimate translation, while the target rotates. In *Sylvester*, **IVT** experiences a complete failure when the target suddenly rotates toward the camera (rapid appearance change). **MIL** follows the target throughout the entire sequence, but at times the lighting and appearance changes (caused by out-of-plane rotations) move the tracking window partially off-target. **MS** also tracks the target throughout the sequence, but allows its target window to grow gradually too large due to a relatively unstructured background and no notion of target spatial organization. **SOE** performs best due to the robustness of its features to illumination changes and their ability to capitalize on motion information when appearance varies rapidly. In *Tiger 2*, **SOE** struggles somewhat relative to **MIL**. Here, the small target combined with rapid motion makes it difficult for the employed coarse-to-fine, gradient-based motion estimator to obtain accurate updates. These challenges make this sequence favorable to trackers that make use of a "spotting approach" (e.g., **MIL**). The result for **SOE** is that it lags behind during the fastest motions; although, it "catches-up" throughout. With *Ming*, **SOE** and **IVT** provide accurate tracks that are qualitatively very similar. **MIL** cannot handle the large scale changes that the target undergoes throughout this video and as such, often ends up only tracking a fraction of the target. **MS** again follows the target but with a tracking window that grows too large. 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 of the **MIL** tracking windows are following the same individual. On the other hand, **MS** and **IVT** cannot surmount the full occlusions caused by the foreground pillar. Only **SOE**'s feature representation, which encompasses target dynamics and spatial organization, is capable of distinguishing between the targets and tracking them both to the conclusion of the video.

In comparing the performance of the proposed **SOE** to the previous approach that made use of spatiotemporal oriented energy features for tracking, **MS**, the benefits of maintaining spatial organization (as provided by **SOE**, but not **MS**) are well documented. **MS** shows a tendency to drift onto non-target locations that share similar feature characteristics with the target when they are collapsed across support regions (e.g., occluding book in *Occluded Face 2*, backgrounds in *Ming* and *Sylvester*). In contrast, **SOE** does not exhibit these problems, as it maintains the spatial organization of the features via its pixelwise representation and the targets are distinguished from the non-target locations on that basis.

Figure 5 shows quantitative error plots for the trackers considered. Since **MIL** is stochastic, it was run 5 times and its errors averaged [3]. Ground truth for *Occluded Face 2*, *Sylvester* and *Tiger 2* were available previously [3]; ground truth was manually obtained for the *Ming* and *Pop Machines* videos. The plots largely corroborate the points that were observed qualitatively. For instance, the minor failure of **MIL** near the end of *Occluded Face 2* is indicated by the rapid increase in error. Also in *Occluded Face 2*, a transient increase in error (Frames 400 — 600) can be observed for **SOE** and **IVT** as they accurately track the target's rotation whereas the ground truth is provided more coarsely as target translation [3]. Similarly, the complete failure of **IVT** near frame 700 of *Sylvester* is readily seen. Also, the tendency of **SOE** to lag and recover in *Tiger 2* is captured in the up/down trace of its error curve, even as it remains generally below that of **IVT** and **MS**, albeit above **MIL**. For *Ming*, the excellent tracks provided by **SOE** and **IVT** are visible. It can also be seen that **MS** experiences a sudden failure as it partially moves off the target near the beginning of the sequence. However, **MS** eventually re-centers itself and provides centers of mass comparable to **MIL**. In the plots for *Pop Machines*, the upward ramps for **IVT**, **MIL**, and **MS** show how the errors slowly increase when the tracking windows fall off of a target as it continues to move progressively away. In contrast, **SOE** enjoys a relatively low error throughout for both targets.

To summarize the quantitative plots in Fig. 5, the center of mass pixel distance error was averaged across all frames, yielding the summary statistics in Table 2. Although the proposed **SOE** does not attain the best performance for every video, it is best in three cases (with one tie) and second best in the remaining two cases (trailing the best by only 3 pixels in one case). **IVT** also scores two top places, in one case tied with, and in the other only slightly better than **SOE**. Further, all trackers except **SOE** experience at least one complete failure where

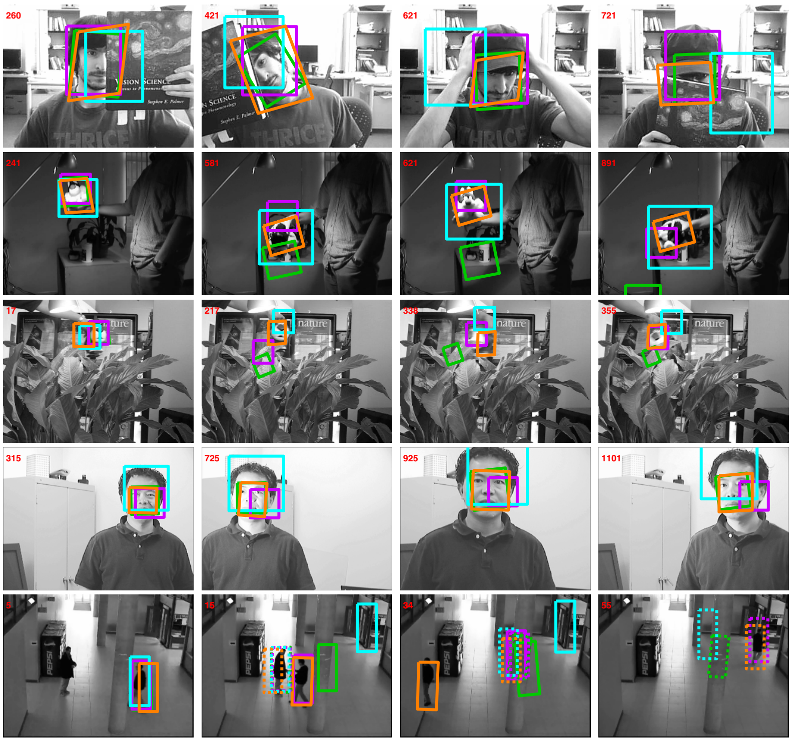

Fig. 4. Comparison to alternative trackers. Top-to-bottom as in Fig. 3. Orange, green, purple, and teal show results for proposed **SOE**, **IVT**, **MIL**, and **MS** trackers, resp.

the tracking window falls off target and does not re-establish a track before the end of the sequence. Overall, these results argue for the superior performance of **SOE** in comparison to the alternatives considered.

4 Discussion and Summary

The main contribution of the presented approach to visual tracking is the introduction of a novel target representation in terms of pixelwise spatiotemporal oriented energies. This representation uniformly captures both the spatial and temporal characteristics of a target with robustness to illumination to yield an uncommonly rich feature set, supporting tracking through appearance and illumination changes, erratic motion, complicated backgrounds and occlusions. A limitation of the current approach is its lack of explicit modeling of background motion, e.g., as encountered with an active camera, which may make the temporal components of the target features less distinctive compared to the background. In future work, various approaches can improve the system in this regard (e.g., background stabilization [8] and automatic selection of a subset of spatiotemporal features distinguishing the target vs. the background [11]).

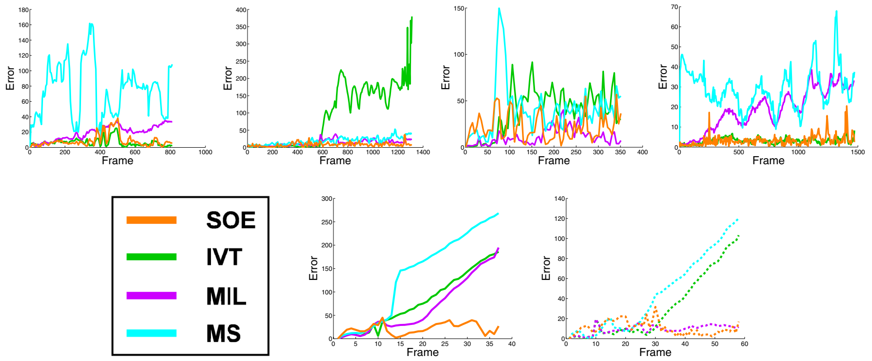

Fig. 5. Quantitative results for Experiment 3. Each plot shows the Euclidean pixel error between the ground truth and tracker center of mass. Row 1, left-to-right, results for *Occluded Face 2*, *Sylvester*, *Tiger 2*, and *Ming*. Row 2, left-to-right, *Pop Machines* target 1 (starting on right) and target 2 (starting on left).

Table 2. Summary of quantitative results. Values listed are pixel distance errors for the center of mass points. Green and red show best and second best performance, resp.

Algorithm	Occluded Face 2	Sylvester	Tiger 2	Ming	Pop Machines
SOE (proposed)	9	8	22	3	13
IVT	6	92	39	3	49
MIL	19	13	11	19	26
MS	75	19	40	29	76

The proposed approach has been realized in a software system for visual tracking that uses robust, parametric motion estimation to capture frame-to-frame target motion. Evaluation of the system on a realistic set of videos confirms the approach's ability to surmount significant tracking challenges (multiple targets, illumination and appearance variation, fast/erratic motion, clutter and occlusion) relative to a variety of alternative state-of-the-art trackers.

References

1. `http://www.cse.yorku.ca/vision/research/oriented-energy-tracking`
2. Adelson, E., Bergen, J.: Spatiotemporal energy models for the perception of motion. JOSA A 2(2), 284–299 (1985)
3. Babenko, B., Yang, M., Belongie, S.: Visual tracking with online multiple instance learning. In: CVPR, pp. 983–990 (2009)
4. Bergen, J., Anandan, P., Hanna, K., Hingorani, R.: Hierarchical model-based motion estimation. In: Sandini, G. (ed.) ECCV 1992. LNCS, vol. 588, pp. 237–252. Springer, Heidelberg (1992)
5. Birchfield, S., Rangarajan, S.: Spatiograms versus histograms for region-based tracking. In: CVPR, vol. 2, pp. 1158–1163 (2005)
6. Black, M., Anandan, P.: The robust estimation of multiple motions: parametric and piecewise-smooth flow fields. CVIU 63(1), 75–104 (1996)

7. Bogomolov, Y., Dror, G., Lapchev, S., Rivlin, E., Rudzsky, M.: Classification of moving targets based on motion and appearance. In: BMVC, pp. 142–149 (2003)
8. Burt, P., Bergen, J., Hingorani, R., Kolczynski, R., Lee, W., Leung, A., Lubin, J., Shvayster, H.: Object tracking with a moving camera. In: Motion Wkshp, pp. 2–12 (1989)
9. Cannons, K.: A review of visual tracking. Technical Report CSE-2008-07, York University, Department of Computer Science and Engineering (2008)
10. Cannons, K., Wildes, R.: Spatiotemporal oriented energy features for visual tracking. In: Yagi, Y., Kang, S.B., Kweon, I.S., Zha, H. (eds.) ACCV 2007, Part I. LNCS, vol. 4843, pp. 532–543. Springer, Heidelberg (2007)
11. Comaniciu, D., Ramesh, V., Meer, P.: Kernel-based object tracking. PAMI 25(5), 564–575 (2003)
12. Derpanis, K., Sizintsev, M., Cannons, K., Wildes, R.: Efficient action spotting based on a spacetime oriented structure representation. In: CVPR (2010)
13. Elgammal, A., Duraiswami, R., Davis, L.: Probabilistic tracking in joint feature-spatial spaces. In: CVPR, pp. 781–788 (2003)
14. Freeman, W., Adelson, E.: The design and use of steerable filters. PAMI 13(9), 891–906 (1991)
15. Granlund, G., Knutsson, H.: Signal Processing for Computer Vision. Kluwer, Dordrecht (1995)
16. Hager, G., Dewan, M., Stewart, C.: Multiple kernel tracking with SSD. In: CVPR, vol. 1, pp. 790–797 (2004)
17. Horn, B.: Robot Vision. MIT Press, Cambridge (1986)
18. Huber, P.: Robust Statistical Procedures. SIAM Press, Philadelphia (1977)
19. Irani, M., Rousso, B., Peleg, S.: Computing occluding and transparent motions. IJCV 12(1), 5–16 (1994)
20. Jepson, A., Fleet, D., El-Maraghi, T.: Robust on-line appearance models for visual tracking. PAMI 25(10), 1296–1311 (2003)
21. Lucas, B., Kanade, T.: An iterative image registration technique with application to stereo vision. In: DARPA IUW, pp. 121–130 (1981)
22. Meyer, F., Bouthemy, P.: Region-based tracking using affine motion models in long image sequences. CVGIP: Image Understanding 60(2), 119–140 (1994)
23. Ross, D., Lim, J., Lin, R., Yang, M.: Incremental learning for robust visual tracking. IJCV 77, 125–141 (2008)
24. Sato, K., Aggarwal, J.: Temporal spatio-velocity transformation and its applications to tracking and interaction. CVIU 96(2), 100–128 (2004)
25. Shi, J., Tomasi, C.: Good features to track. In: CVPR, pp. 593–600 (1994)
26. Sizintsev, M., Wildes, R.: Spatiotemporal stereo via spatiotemporal quadric element (stequel) matching. In: CVPR, pp. 493–500 (2009)
27. Stauffer, C., Grimson, W.: Learning patterns of activity using real-time tracking. PAMI 22(8), 747–757 (2000)
28. Takala, V., Pietikainen, M.: Multi-object tracking using color, texture and motion. In: ICCV (2007)
29. Wildes, R., Bergen, J.: Qualitative spatiotemporal analysis using an oriented energy representation. In: Vernon, D. (ed.) ECCV 2000. LNCS, vol. 1843, pp. 768–784. Springer, Heidelberg (2000)
30. Wren, C., Azarbayejani, A., Darrell, T., Pentland, A.: Pfinder: Real-time tracking of the human body. PAMI 19(7), 780–785 (1997)
31. Zaharescu, A., Wildes, R.: Anomalous behaviour detection using spatiotemporal oriented energies, subset inclusion histogram comparison and event-driven processing. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part I. LNCS, vol. 6311, pp. 563–576. Springer, Heidelberg (2010)