# Visual Tracking using a Pixelwise Spatiotemporal Oriented Energy Representation


Kevin Cannons

Supervisor: Richard P. Wildes

{kcannons, wildes}@cse.yorku.ca

Department of Computer Science and Engineering, York University


# INTRODUCTION

## Motivation

- There are several direct applications of "following a target" (e.g., surveillance, active camera systems).
- Many computer vision problems are beginning to rely on visual trackers as an initial stage of processing (e.g., activity recognition, object recognition).
- Key challenges for visual trackers include: illumination effects, scene clutter, and sudden changes in target appearance or velocity.
- It is proposed that the choice of representation is key to meeting the above challenges.

**GOAL:** To identify a rich, pixelwise representation of a target that models both its spatial and dynamic properties in a uniform fashion and can be instantiated effectively in a tracker.

## Contributions

(i) A novel oriented energy representation that retains the spatial organization of the target is developed for visual tracking. 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 as well or better than several state-of-the-art algorithms during extensive qualitative and quantitative comparisons.

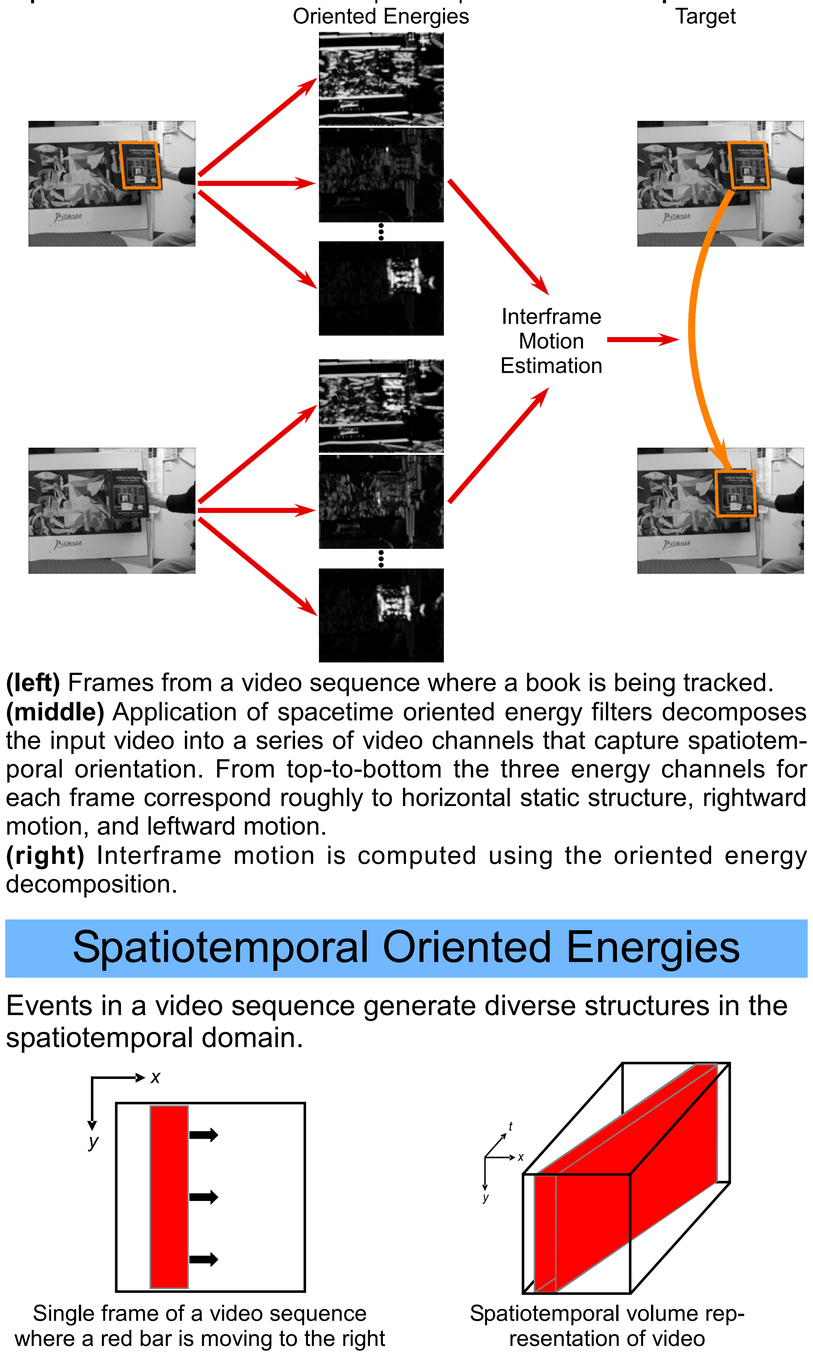
# TECHNICAL APPROACH

## Overview

**Input:** Video Frames

**Features:** Spatiotemporal Oriented Energies

**Output:** Tracked Target




**(left)** Frames from a video sequence where a book is being tracked.
**(middle)** Application of spacetime oriented energy filters decomposes the input video into a series of video channels that capture spatiotemporal orientation. From top-to-bottom the three energy channels for each frame correspond roughly to horizontal static structure, rightward motion, and leftward motion.
**(right)** Interframe motion is computed using the oriented energy decomposition.

## Spatiotemporal Oriented Energies

Events in a video sequence generate diverse structures in the spatiotemporal domain.




Single frame of a video sequence where a red bar is moving to the right




Spatiotemporal volume representation of video

In the figure above, the rightward motion appears as a diagonal structure extending from the front left to the back right of the cube in the spatiotemporal domain. Structures of arbitrary orientation can be generated in the spacetime domain by the various static structures and motions that can occur in a video.

One method of capturing the spatiotemporal characteristics of a video sequence is through the use of oriented energies [1, 2]. These energies are derived using the filter responses of orientation selective bandpass filters when they are convolved with a spatiotemporal volume.

The desired oriented energies are realized using broadly tuned 3D Gaussian second derivative filters, $G_2(\theta,\gamma)$, and their corresponding Hilbert transforms, $H_2(\theta,\gamma)$, where $\theta$ specifies the 3D direction of the filter axis of symmetry, and $\gamma$ indicates the scale within a Gaussian pyramid formulation [3]. Hence, an initial measure of local energy can be computed according to

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

where $\mathbf{x} = (x, y, t)$ are spatiotemporal image coordinates, $I$ is an image, and $*$ denotes the convolution operator.

To obtain a purer measure of the relative contribution of orientations, irrespective of image contrast, pixelwise normalization is performed,

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

where $\epsilon$ is a constant introduced to avoid instabilities when the overall energy content is small.

For this work, energies were computed at 10 orientations, as they span the space of 3D orientations for the highest order filters that were used (i.e., $H_2$). Energies were computed at a single scale, corresponding to direct application of the oriented filters to the input imagery.

## Robust Motion Estimation

Tracking using a pixelwise template approach consists of matching the template to the current frame to estimate the interframe motion of the target. Here, both the template and the image frame are represented in terms of oriented energies.

The optical flow constraint equation [4] is used to formulate a measure of match between feature measurements (oriented energies) that are aligned by a parametric motion model

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

where $\nabla^\top \hat{E} = \left(\hat{E}_x, \hat{E}_y\right)$ are the first-order spatial derivatives of the image energy measurements for a specific scale and orientation, $\mathbf{u} = (u, v)^\top$ 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 parameters, $\mathbf{a}$, are estimated by minimizing the error in the optical flow constraint equation, summed over the target support. Significantly, in the present approach the target representation spans multiple feature channels (orientations and scales) of spatiotemporal oriented energies, leading to

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

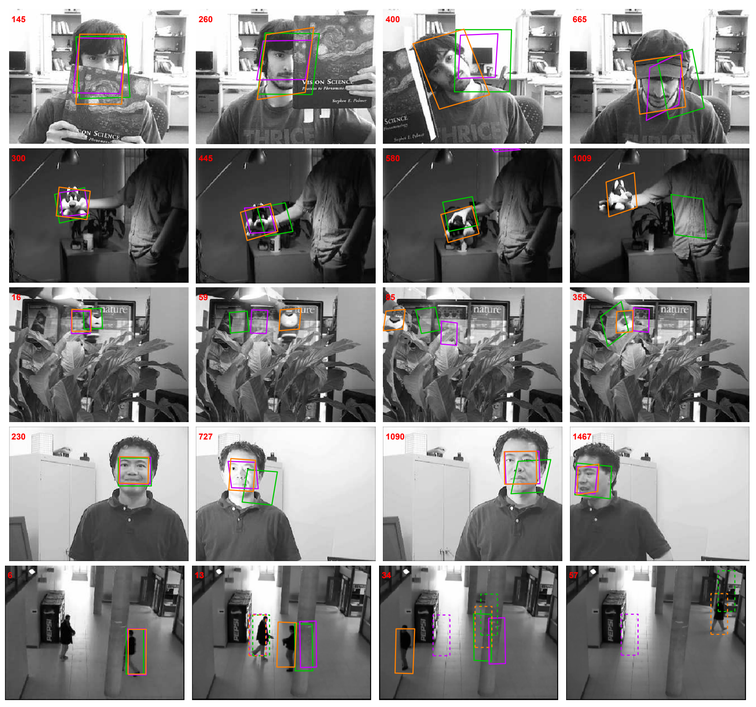
where $\rho$ is the Geman-McClure error metric with width $\sigma$.

# EMPIRICAL EVALUATION

## Feature Set Comparison

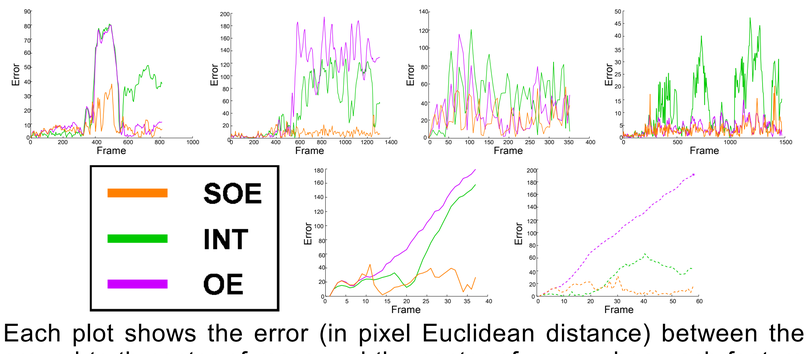
This experiment provides a direct comparison between the proposed spatiotemporal oriented energy features and two alternatives. Three tracking systems were considered, the single difference between them being the feature set. The first tracker, **SOE**, used the proposed spatiotemporal oriented energy features (10 spacetime orientations) while the second tracker, **INT**, simply employed pixelwise raw image intensities. The third tracker, **OE**, utilized a purely spatial oriented energy feature representation. For **OE**, the energies were computed at four spatial orientations within the image plane.

### Qualitative Results



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* videos. Orange, purple, and green boxes are for **SOE**, **OE**, and **INT** trackers, resp.

### Quantitative Results




Each plot shows the error (in pixel Euclidean distance) between the ground truth center of mass and the center of mass when each feature representation was used for tracking. 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).

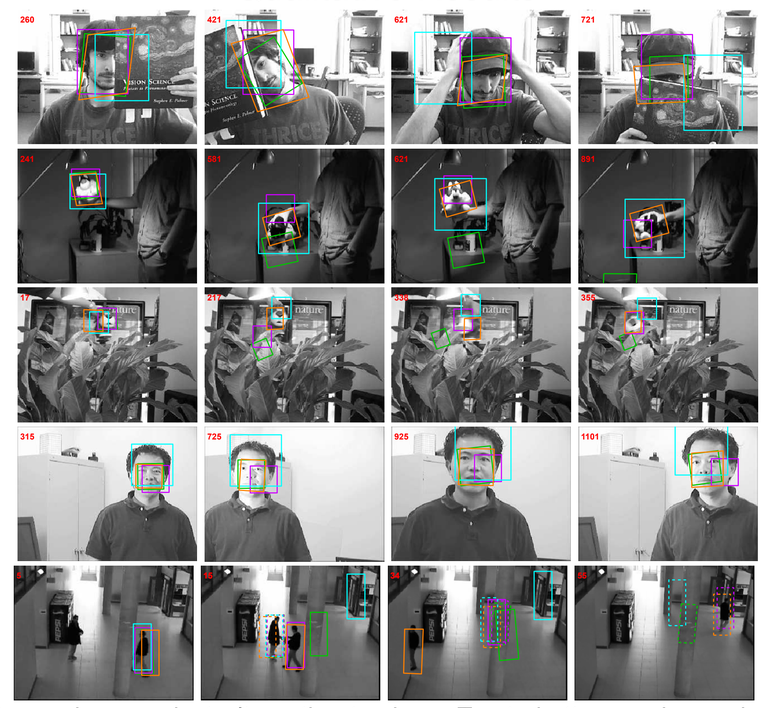
Summary of quantitative results for feature set experiment. Values listed are pixel distance errors for the center of mass points averaged over all frames. Green and red show best and second best performance, resp.

| Algorithm | Occluded Face 2 | Sylvester | Tiger 2 | Ming | Pop Machines |
|---|---|---|---|---|---|
| SOE (proposed) | 9 | 8 | 22 | 3 | 13 |
| INT | 27 | 51 | 46 | 13 | 39 |
| OE | 18 | 81 | 33 | 4 | 97 |

## Comparison Against Strong Trackers

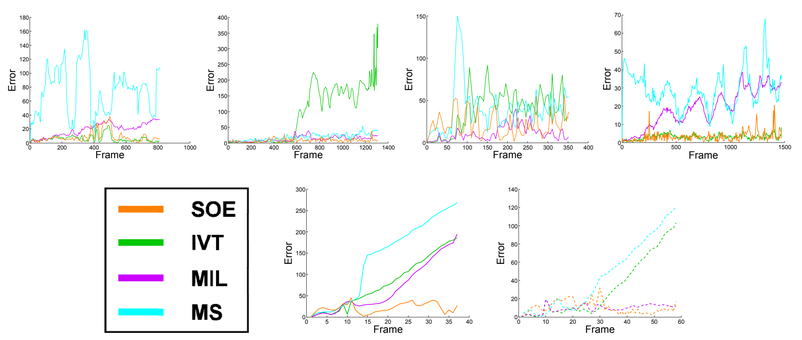
This experiment compares the performance of the proposed overall spatiotemporal oriented energy tracker, **SOE**, against several strong trackers. The specific trackers considered are the multiple instance learning tracker (**MIL**) [5], the incremental visual tracker (**IVT**) [6], and a tracker that uses a similar oriented energy representation, but that is spatially collapsed across target support to fit within the mean shift framework (**MS**) [7].

### Qualitative Results



Comparison against alternative trackers. Top-to-bottom as in previous qualitative results figure. Orange, green, purple, and teal boxes show results for proposed **SOE**, **IVT**, **MIL**, and **MS** trackers, resp.

### Quantitative Results




Each plot shows the error (in pixel Euclidean distance) between the ground truth center of mass and the center of mass provided by each tracker. Left-to-right, as in previous quantitative results figure.

Summary of quantitative results for strong tracker experiment. Values listed are pixel distance errors for the center of mass points averaged over all frames. 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 |

# SUMMARY

(i) A novel representation in terms of pixelwise spatiotemporal oriented energies was derived for the problem of tracking.

(ii) The representation was shown to outperform other commonly-used features and the resulting tracker provided superior performance relative to several state-of-the-art trackers.

(iii) Superior performance was attained because of the richness of the representation (uniformly capturing spatial and temporal characteristics) and its robustness to illumination.

## References


[1] Adelson, E., Bergen, J.: Spatiotemporal energy models for the perception of motion. JOSA 2(2), (1985)
[2] Wildes, R., Bergen, J.: Qualitative spatiotemporal analysis using an oriented energy representation. ECCV (2000)
[3] Freeman, W., Adelson, E.: The design and use of steerable Filters. PAMI 13 (1991)
[4] Horn, B.: Robot Vision. MIT Press (1986)
[5] Babenko, B., et al.: Visual tracking with online multiple instance learning. In: CVPR. (2009)
[6] Ross, D.A., et al.: Incremental learning for robust visual tracking. IJCV 77 (2008)
[7] Cannons, K., Wildes, R.: Spatiotemporal oriented energy features for visual tracking. In: ACCV. (2007)