Spatiotemporal Oriented Energy Features for Visual Tracking

Kevin Cannons and Richard Wildes

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

Abstract. This paper presents a novel feature set for visual tracking that is derived from "oriented energies". More specifically, energy measures are used to capture a target's multiscale orientation structure across both space and time, yielding a rich description of its spatiotemporal characteristics. To illustrate utility with respect to a particular tracking mechanism, we show how to instantiate oriented energy features efficiently within the mean shift estimator. Empirical evaluations of the resulting algorithm illustrate that it excels in certain important situations, such as tracking in clutter with multiple similarly colored objects and environments with changing illumination. Many trackers fail when presented with these types of challenging video sequences.

1 Introduction

Target tracking is a critically important aspect to a wide range of computer vision applications, including surveillance, smart rooms, and human-computer interfaces. Significant contributions have been made to the field, but no general-purpose tracker has been found that can operate effectively in every real-world setting [1]. Scenarios that are present in realistic sequences and challenge many trackers include changes in illumination, small targets, and significant clutter.

In general, to facilitate accurate tracking, features must be selected that distinguish targets from the background and from one another, even while being robust to photometric and geometric distortions. In response to these requirements, many different proposals have been made; here, representative examples are provided. Perhaps the simplest approach is to make use of image intensity-based templates for feature definition [2,3,4]. To provide robustness to photometric distortions, consideration has been given to discrete features [5,6,7]. To encompass object outlines, methods have emerged that use contours and silhouettes [8,9,10]. Other features (e.g., color, texture) have been derived on a more regional basis [11,12,13]. Recovered motion also has been used in feature definitions [14,15,16].

Limited attention has been given to the integrated analysis of both the spatial and temporal domains when considering features for visual tracking. Potential benefits of a more integrated approach include the ability to combine static and dynamic target information in a natural fashion as well as simplicity of

Y. Yagi et al. (Eds.): ACCV 2007, Part I, LNCS 4843, pp. 532–543, 2007.

design and implementation. In response to this observation, the present paper documents a novel feature set for visual tracking that uses energy measures to capture a target's multiscale, spatiotemporal orientation structure. A considerable body of research has emerged on the use of orientation selective filters in the spatiotemporal domain for the purpose of analyzing motion [17,18,19]. However, it appears that no previous research has explored the use of multiscale, spatiotemporal oriented energies that uniformly encompass space and time as the basis for defining features in the service of visual tracking.

To illustrate the use of the proposed oriented energy feature set, we make use of the mean shift tracking paradigm [13,20,21,22], a framework upon which these features readily map. Although, the energy features are also applicable to alternative paradigms, e.g., those that preserve within target spatial relationships, as the oriented energies are calculated locally.

In light of previous research, the main contributions of this paper are as follows. (1) A novel oriented energy feature set is defined for visual tracking. This representation captures the spatiotemporal characteristics of a target in an integrated, compact fashion. (2) Oriented energy features are instantiated with respect to the mean shift estimator. (3) The performance of the resulting system is documented both qualitatively and quantitatively. Our algorithm outperforms a color-based mean shift implementation in three common, real-world situations: substantial clutter; multiple targets with similar color; and illumination changes.

2 Technical Approach

2.1 Oriented Energy Features

Oriented Energy Computation. Events in a video sequence will generate diverse structures in the spatiotemporal domain. For instance, a textured, stationary object produces a much different signature in image space-time than if the same object were moving. One method of capturing the spatiotemporal characteristics of a video sequence is through the use of oriented energies [17]. These energies are derived using the filter responses of orientation selective bandpass filters when they are convolved with the spatiotemporal volume produced by a video stream. Responses of filters that are oriented parallel to the image plane are indicative of the spatial pattern of observed surfaces and objects (e.g., spatial texture); whereas, orientations that extend into the temporal dimension capture dynamic aspects (e.g., velocity and flicker).

The basis of our approach is that energies computed at orientations which span the space-time domain can provide a rich description of a target for visual tracking. Here, multiscale processing is also important, as coarse scales capture gross spatial pattern and overall target motion while finer scales capture detailed spatial pattern and motion of individual parts (e.g., limbs). With regard to dynamic aspects, simple motion is captured (orientation along a single spatiotemporal diagonal) as well as more complex phenomena, e.g., multiple juxtaposed motions as limbs cross (multiple orientations in a spatiotemporal region). By encompassing both spatial and temporal target characteristics in an integrated fashion,

tracking is supported in the presence of significant clutter. Further, as detailed below, such representations can be made invariant to local image contrast to support tracking throughout substantial illumination changes.

For this work, filtering was performed using broadly tuned, steerable, separable filters based on the second derivative of a Gaussian, G_2, and their corresponding Hilbert transforms, H_2 [23], with responses pointwise rectified (squared) and summed. Filtering was executed across $\theta = (\eta, \xi)$ 3D orientations (η, ξ specifying polar angles) and σ scales using a Gaussian pyramid formulation. Hence, a measure of local energy, e, can be computed according to

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

where $\mathbf{x} = (x, y, t)$ corresponds to spatiotemporal image coordinates, I is the image sequence, and $*$ denotes convolution. This initial measure of local energy is dependent on image contrast. To attain a purer measure of the relative contribution of the orientations irrespective of contrast, (1) is normalized as

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

where ϵ is a bias term to avoid instabilities when the energy content is small and the summations in the denominator cover all scale and orientation combinations. (In this paper, our convention is to superscript variables of summation with $\tilde{.}$)

For illustrative purposes, Fig. 1 displays a subset of the energies that are computed for a single frame of a MERL traffic sequence [24]. Here, there is a white car moving to the left near the center of the frame. Notice how the energy channel that is tuned for leftward motion is very effective at distinguishing this car from the static background. Consideration of the channel tuned for horizontal structure shows how it captures the overall orientation structure of the white car. In contrast, while the channel tuned for vertical textures captures the outline of the crosswalks, it shows 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.

Given that the tracking problem is being considered, the goal is to locate the target's position as precisely as possible. However, as seen in Fig. 1, the energies computed at coarser scales are diffuse due to the downsampling/upsampling that is employed in pyramid processing. Coarse energies are important because they provide information regarding the target's gross shape and motion, but a method is required to improve their localization for accurate tracking. To that end, a set of weights are applied to the normalized energies of (2) according to

$$\hat{E}\left(\mathbf{x};\theta,\sigma\right) = \hat{e}\left(\mathbf{x};\theta,\sigma\right) b\left(\mathbf{x};\theta\right), \tag{3}$$

where b are pixel-wise 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, T_θ, according to

$$b\left(\mathbf{x};\theta\right) = \sum_{\tilde{\sigma}} \hat{e}\left(\mathbf{x};\theta,\tilde{\sigma}\right) > T_\theta. \tag{4}$$

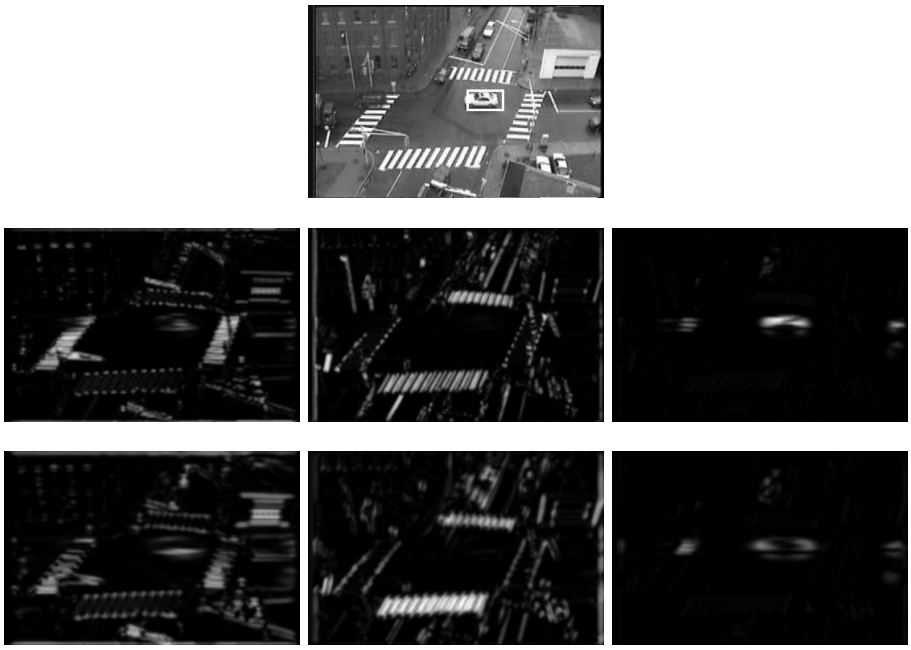

Fig. 1. Frame 29 of the MERL traffic video sequence with select corresponding energy channels. Finer and coarser scales are shown in rows two and three, resp. From left to right, the energy channels roughly correspond to horizontal structure, vertical structure, and leftward motion.

When computing the weights, 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, we avoid being prejudiced toward any particular type of oriented structure (e.g., static vs. dynamic).

Two significant advantages of the proposed oriented energy feature set must be further highlighted. First, normalized energy, as defined by (1) and (2), captures local spatiotemporal structure at a particular orientation and scale with a degree of robustness to scene illumination: By virtue of the bandpass filtering, (1), invariance will be had to changes that are manifest in the image as additive offsets to image brightness; by virtue of the normalization, (2), invariance will be had to changes that are manifest in the image as multiplicative offsets. Second, the calculation of the defined normalized oriented energies requires nothing more than 3D separable convolution and pointwise nonlinear operations, and is thereby amenable to compact, efficient implementation [25].

Histogram Representation. As defined, oriented energies provide local characterization of image structure. Therefore, the energy measurements could be used to provide pointwise descriptors for target tracking (e.g., in conjunction

with spatial template-based matching). Alternatively, the pointwise measurements can be aggregated over target support to provide region-based descriptors (e.g., in conjunction with mean shift tracking). Here, we pursue the second option and demonstrate the efficacy of the features as regional descriptors.

With an eye to mean shift tracking, we collapse the spatial information in our initial energy measurements and represent the target as a histogram. Each histogram bin corresponds to the weighted energy content of the target at a particular scale and orientation. Specifically, the template histogram that defines the target in the first frame is given by

$$\hat{q}_u = C \sum_{i=1}^{n} k\left(\|\mathbf{x}_i^*\|^2\right) \hat{E}\left(\mathbf{x}_i^*; \phi_u\right), \tag{5}$$

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 single 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.

When tracking a target, it may be necessary to evaluate several target candidates for the current frame. Candidate histograms are defined as

$$\hat{p}_u\left(\mathbf{y}\right) = C_h \sum_{i=1}^{n_h} k\left(\left|\left|\frac{\mathbf{y} - \mathbf{x}_i^*}{h}\right|\right|^2\right) \hat{E}\left(\mathbf{x}_i^*; \phi_u\right), \tag{6}$$

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 oriented energies for the highest frequency structures have the strongest response, as the target is fairly small and dominated by relatively finer scale structure.

2.2 Oriented Energy Features in the Mean Shift Framework

Target Position Estimation. Under the mean shift framework, tracking an object involves locating the candidate position in the current frame that produces the histogram that is most similar to the template. Thus, a measure of similarity between two histograms is required. For histogram comparisons we utilize the Bhattacharyya coefficient, the sample estimate of which can be computed using

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

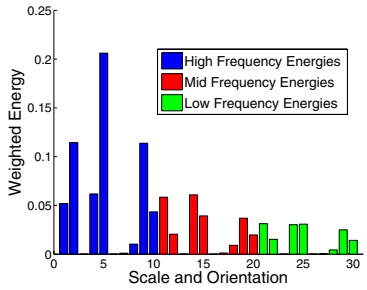

Fig. 2. Oriented energy histogram for the target region in Fig. 1

where $\hat{\mathbf{p}}(\mathbf{y})$ and $\hat{\mathbf{q}}$ are histograms with m bins apiece. Due to the definition of the Bhattacharyya coefficient, in order to minimize the distance between two histograms, (7) must be maximized with respect to the target position, $\mathbf{y}$.

The Bhattacharyya coefficient can be maximized via mean shift iterations [20]. The specific mean shift vector that can be used for this maximization is

$$\hat{\mathbf{y}}_1 = \left[\frac{\sum_{i=1}^{n_h} \mathbf{x}_i^* w_i g\left(\left\| \frac{\hat{\mathbf{y}}_0 - \mathbf{x}_i^*}{h} \right\|^2 \right)}{\sum_{i=1}^{n_h} w_i g\left(\left\| \frac{\hat{\mathbf{y}}_0 - \mathbf{x}_i^*}{h} \right\|^2 \right)} \right], \text{ where } \quad w_i = \sum_{u=1}^{m} \hat{E}\left(\mathbf{x}_i^*; \phi_u\right) \sqrt{\frac{\hat{q}_u}{\hat{p}_u\left(\hat{\mathbf{y}}_0\right)}}, \tag{8}$$

$g(x) = k'(x)$ is the derivative of the tracking kernel profile, k, with respect to x, and $\hat{\mathbf{y}}_0$ is the current target position. The Epanechnikov kernel has been shown to be effective [20] and is the most commonly used kernel for mean shift tracking.

Thus, the position of the target in the current frame is estimated as follows. Starting from the target's position in the previous frame, the mean shift vector is computed and the target candidate is moved to the position indicated by the mean shift vector. These steps are repeated until convergence has been reached or a fixed number of iterations have been executed.

Template and Scale Updates. When tracking an object through a long video sequence, it is common that its characteristics will change. To combat the changes a target may incur over time (e.g., due to alterations in velocity or rotation), our tracker includes a simple template update mechanism defined as

$$\hat{\mathbf{q}}^{i+1} = \alpha\pi\hat{\mathbf{q}}^i + (1-\alpha)(1-\pi)\,\hat{\mathbf{p}}(\mathbf{y}_i), \tag{9}$$

where α is a weighting factor to control the speed of the updates, $\hat{\mathbf{q}}^i$ is the template at frame i, and $\pi = \rho\left[\hat{\mathbf{p}}(\mathbf{y}_i), \hat{\mathbf{q}}^i\right]$ is the Bhattacharyya coefficient between the current template and the optimal candidate found in the i^{th} frame. Empirically, α was set to 0.85. Following each application of (9), the resulting template is renormalized and thereby remains consistent with our overall formulation. Owing to dependence on the Bhattacharyya coefficient, the template update rule

indicates that if the template and the optimal candidate are well-matched, the update to the template will be minimal.

The size of a target may change during a video sequence as well. Although there are more effective methods of dealing with changes of object scale in the mean shift framework [21,22], in the current implementation we employ a simple approach, similar to that taken in [20]. In particular, our system performs mean shift optimization three times per frame using three different bandwidth values, h. Unless stated otherwise, h values of $\pm 5\%$ are used. We obtain the new bandwidth, h_{new}, by combining the best of the three bandwidths evaluated at the current frame, h_{opt}, with the previous target size, h_{prev}, according to

$$h_{\text{new}} = \gamma h_{\text{opt}} + (1 - \gamma)\, h_{\text{prev}}. \tag{10}$$

Empirically, we set $\gamma = 0.15$.

3 Empirical Evaluation

The performance of the oriented energy-based mean shift tracker has been evaluated on an illustrative set of test sequences. For comparative purposes, a mean shift tracker based on RGB color space was also developed and tested. Other than the use of different histograms, the two trackers were identical. The color-based tracker was implemented in a similar manner to [20], whereby each color channel was quantized into 16 levels (yielding a histogram with 16^3 bins). In our current implementation of the energy-based tracker, energies were computed at 3 scales with 10 different spatiotemporal orientations per scale. Hence, the energy-based histograms contained 30 bins. For the oriented energy feature set, 10 orientations were selected because they span the space of 3D orientations for the highest order filters that we use (H_2) [23]; in particular, the selected orientations correspond to the normals to the faces of an icosahedron with antipodal directions counted once, which provides a uniform tessellation of a sphere. For all results in this paper, an Epanechnikov kernel, K, was used. The thresholds for (4) were empirically set as $2.75\times$ the mean energy for each orientation channel. The color and energy-based trackers were hand-initialized with identical target regions in the first frame of each video.

Figure 3 illustrates the effectiveness of oriented energy-based features in dealing with illumination changes. An individual starts walking in a poorly lit area;

Fig. 3. Video sequence ($x \times y \times t = 360 \times 240 \times 60$) of a man walking through shadows. From left to right, frames 4, 18, 31, and 55 are shown. Tracked regions are highlighted with white boxes.

Fig. 4. Video sequence ($x \times y \times t = 360 \times 240 \times 50$) of people walking through a room with similar colored clothing. From left to right, frames 6, 18, 32, and 50 are shown. Tracked regions are highlighted with white boxes.

Fig. 5. MERL traffic video sequence ($x \times y \times t = 368 \times 240 \times 64$) where a white car is tracked as it travels through an intersection. From left to right frames 13, 24, 38, and 58 are shown. Tracked regions are highlighted with white boxes.

then, he travels into and out of the bright region as he walks across the room. Using our proposed feature set, the tracker appeared to be relatively unaffected by the changes in illumination. This robustness arises from the normalization performed in (2). In comparison, our color-based mean shift tracker completely lost track of the target after only a few frames, even when histograms created using normalized RG-space [20] were utilized.

Figure 4 shows a case where two persons with similar colored clothing walk in opposite directions and the individual starting on the right side is being tracked. Despite the full occlusion that occurs for several frames, the tracker using energy features is capable of following the true target throughout the video. The different texture patterns and velocities of the walkers were sufficient cues for the energy-based tracker to achieve success, as the representation spans the spatiotemporal domain. In comparison, our color-based tracker became distracted by the other walker as the individuals have near-identical color distributions.

Figure 5 shows a real-life, grayscale video sequence of a cluttered traffic scene that was obtained from MERL [24] (a portion also used in Fig. 1). As the figure shows, our proposed system experiences some slight difficulty when tracking the vehicle as it passes over the crosswalk (e.g. notice off-centered tracking in frames 13 and 24). This performance decrease occurs because the lack of contrast (essentially uniform white on white) between the car and the crosswalk yields little energy for the involved portions of the car. Nevertheless, the tracker never loses the target; indeed, the frames shown are representative of the worst case performance in this video.

Our feature set was also successfully used when tracking people and vehicles in videos obtained from the PETS2001 dataset [26]. Figure 6 shows an example

Fig. 6. PETS2001 video sequence ($x \times y \times t = 384 \times 288 \times 85$) where a cyclist is being tracked. From left to right frames 18, 32, and 73 are displayed. Tracked regions are highlighted with white boxes.

Fig. 7. Video sequence ($x \times y \times t = 360 \times 240 \times 100$) showing an individual walking in an erratic pattern. From left to right frames 22, 74, 86, and 100 are displayed. Tracked regions are highlighted with white boxes.

of our results on this dataset where a cyclist is tracked. The tracker that utilizes oriented energy features is successful despite the fact that the cyclist is partially occluded by another individual near the beginning of the sequence. The results on this data sequence are impressive given that the video accurately reflects real-world surveillance settings where targets of interest are often small and of low-resolution. In contrast, our implementation of the color-based mean shift tracker drifted off the target after only a few frames.

In Fig. 7 an individual is shown walking erratically, making sudden changes in direction and moving at a wide variety of speeds. Since the oriented energy features encompass both spatial and temporal information, tracking of the target continues throughout each change in velocity. In particular, at instances where the target motion changes radically, the spatially-based components of the representation keep the tracker on target. Subsequently, template updates, (9), incorporate changes to adapt the model for further tracking.

Figure 8 shows footage that one might obtain from overhead surveillance cameras in public areas. The oriented energy-based tracker follows the target of interest even though there are multiple similar walkers with little texture, cast shadows, and complex reflectance effects, as the video was recorded through a window. Using the oriented energy feature set, the target is not lost, even during the partial occlusion. The tracker does lag behind the target for a few frames immediately following the occlusion; however, it ultimately follows the correct person. Indeed, frame 39 is representative of its worst-case performance for this

Fig. 8. Video sequence ($x \times y \times t = 320 \times 240 \times 70$) showing multiple people in motion that are similar in appearance. From left to right frames 9, 31, 39, and 59 are displayed. Tracked regions are highlighted with white boxes.

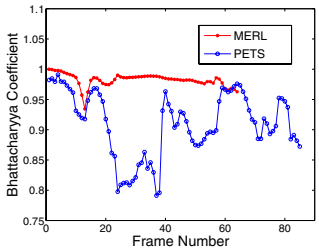

Fig. 9. Bhattacharyya coefficients over the entire video sequence for the MERL and PETS2001 videos

video. In comparison, our color-based implementation was only able to follow the true target for approximately 30 frames.

Quantitative performance analysis was performed for the video sequences that are publicly available — MERL and PETS2001. Specifically, Fig. 9 shows the Bhattacharyya coefficient vs. frame number for these two sequences. The Bhattacharyya coefficient is a measure of the system's confidence in the target found in each frame, with 1 being the largest possible value. For the MERL video, the decreased level of performance at the crosswalks that was qualitatively observed is also indicated quantitatively. In particular, Fig. 9 shows two slight decreases in the Bhattacharyya coefficient at frames 12 and 58 — precisely the frames when the vehicle is passing over the crosswalks. For the PETS video sequence, the significant deviation the Bhattacharyya coefficient experiences is a result of the partial occlusion of the cyclist by the walker (approximately frames 15 - 34). The other, less substantial decreases are a result of the significant background clutter (e.g., parked cars). Also of note is that an average of 3 mean shift iterations were required to reach convergence for these two videos. Twenty iterations, the maximum we allow, was observed only three times.

4 Summary

Spatiotemporal oriented energy features provide a rich, yet compact representation of a target's characteristic structure across both space and time. In particular,

by encompassing a range of orientations and scales, the proposed feature set provides a natural integration of the static (e.g., spatial texture) and dynamic (e.g., motion) aspects of a target. To illustrate their usefulness with respect to a particular tracking mechanism, we provide an instantiation with respect to the mean shift estimator. In our experiments over a wide range of video sequences, the energy-based tracker was considered to perform as well or better than an identical algorithm that used color histograms. Of primary interest in our work were surveillance-inspired video sequences that included challenges such as substantial background clutter, targets that contained similar colors to other objects in the scene, and changes in illumination. Tracking with the use of oriented energy features was shown to be robust to these challenges.

Acknowledgments. Portions of this work were funded by an Ontario Graduate Scholarship to K. Cannons and an NSERC Discovery Grant to R. Wildes.

References

1. Yilmaz, A., Javed, O., Shah, M.: Object tracking: A survey. Comp. Surv. 38(4), 1–45 (2006)
2. Lucas, B., Kanade, T.: An iterative image registration technique with application to stereo vision. In: DARPA IUW, pp. 121–130 (1981)
3. Anandan, P.: A computational framework and an algorithm for the measurement of visual motion. IJCV 2(3), 283–310 (1989)
4. Shi, J., Tomasi, C.: Good features to track. CVPR 1, 593–600 (1994)
5. Sethi, I., Jain, R.: Finding trajectories of feature points in monocular images. PAMI 9(1), 56–73 (1987)
6. Deriche, R., Faugeras, O.: Tracking line segments. IVC 8(4), 261–270 (1991)
7. Rangarajan, K., Shah, M.: Establishing motion correspondence. CVGIP 54(1), 56–73 (1991)
8. Terzopoulos, D., Szeliski, R.: Tracking with kalman snakes. In: Blake, A., Yuille, A. (eds.) Active Vision, pp. 553–556. MIT Press, Cambridge (1992)
9. Isard, M., Blake, A.: Contour tracking by stochastic propagation of conditional density. In: Buxton, B.F., Cipolla, R. (eds.) ECCV 1996. LNCS, vol. 1064, pp. 343–354. Springer, Heidelberg (1996)
10. Haritaoglu, L., Harwood, D., Davis, L.: W4: Real-time surveillance of people and their activities. PAMI 22(8), 809–830 (2000)
11. Birchfield, S.: Elliptic head tracking with intensity gradients and color histograms. CVPR 1, 232–237 (1998)
12. Sigal, L., Sclaroff, S., Athitsos, V.: Estimation and prediction of evolving color distributions for skin segmentation under varying illumination. CVPR 2, 152–159 (2000)
13. Elgammal, A., Duraiswami, R., Davis, L.: Probabilistic tracking in joint feature-spatial spaces. CVPR 1, 781–788 (2003)
14. Bolgomolov, Y., Dror, G., Lapchev, S., Rivlin, E., Rudzsky, M.: Classification of moving targets based on motion and appearance. In: BMVC, pp. 142–149 (2003)
15. Cremers, D., Schnorr, C.: Statistical shape knowledge in variational motion segmentation. IVC 21(1), 77–86 (2003)

16. Sato, K., Aggarwal, J.: Temporal spatio-velocity transformation and its application to tracking and interaction. CVIU 96(2), 100–128 (2004)
17. Adelson, E., Bergen, J.: Spatiotemporal energy models for the perception of motion. JOSA 2(2), 284–299 (1985)
18. Heeger, D.: Optical flow from spatiotemporal filters. IJCV 1(4), 297–302 (1988)
19. Enzweiler, M., Wildes, R., Herpers, R.: Unified target detection and tracking using motion coherence. Wrkshp. Motion & Video Comp. 2, 66–71 (2005)
20. Comaniciu, D., Ramesh, V., Meer, P.: Kernel-based object tracking. IEEE PAMI 25(5), 564–575 (2003)
21. Collins, R.: Mean-shift blob tracking through scale space. CVPR 2, 234–240 (2003)
22. Zivkovic, Z., Krose, B.: An EM-like algorithm for color-histogram tracking. CVPR 1, 798–803 (2004)
23. Freeman, W., Adelson, E.: The design and use of steerable filters. IEEE PAMI 13(9), 891–906 (1991)
24. Brand, M., Kettnaker, V.: Discovery and segmentation of activities in video. IEEE PAMI 22(8), 844–851 (2000)
25. Derpanis, K., Gryn, J.: Three-dimensional nth derivative of Gaussian separable steerable filters. ICIP 3, 553–556 (2005)
26. PETS (2006), `http://peipa.essex.ac.uk/ipa/pix/pets/`