Spatiotemporal Oriented Energy Features for Visual Tracking

Kevin Cannons

Richard P. Wildes

Technical Report CSE-2007-02

May 22, 2007

Department of Computer Science and Engineering

4700 Keele Street North York, Ontario M3J 1P3 Canada

Spatiotemporal Oriented Energy Features for Visual Tracking

Kevin Cannons
Richard P. Wildes

Department of Computer Science and Engineering
and the Centre for Vision Research
York University
Toronto, Ontario M3J 1P3
Canada

May 22, 2007

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. Color-based trackers often fail when presented with these types of challenging video sequences.

Contents

Chapter 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 (see, e.g., [1] for a general review). Common challenges that are present in realistic sequences include large changes in illumination, small targets and significant clutter. Accordingly, tracking is still considered to be an open problem in computer vision.

To facilitate accurate tracking, features must be selected that distinguish targets from the background and from one another. Further, the features must be robust to photometric and geometric image distortions. Moreover, because tracking applications often require rapid updates, features of interest must lend themselves to efficient extraction. In response to these requirements, many different proposals have been made and space prohibits an exhaustive survey; here, representative examples are provided. Perhaps the simplest approach is to make use of image intensity-based templates for feature definition [2, 3, 4]. Along this line, analysis has been developed to select template windows that will yield most accurate tracks [5]. To provide robustness to photometric distortions, consideration has been given to discete features, e.g., edges, lines and corners [6, 7, 8, 9, 10]. To encompass object outlines, methods have emerged that use contours and silhouettes [11, 12, 13]. Other features are derived on a more regional basis, e.g., color, texture and their combination [14, 15, 16, 17, 18]. Recovered motion also has been incorporated in feature definitions on its own and in combination with appearance [19, 20, 21]. Yet another class of approaches rely on learning

methods to discover relevant features with respect to a training set [22, 23].

Across the wide range of features considered for visual tracking, limited attention has been given to the integrated analysis of both the spatial and temporal domains. The 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 design and implementation. In response to this observation, the present paper documents a novel feature set for visual tracking that is based on energy measures for capturing a target's multiscale, spatiotemporal orientation structure. It is shown that the resulting representation yields rich, yet compact target descriptions that naturally integrate both temporal and spatial characteristics in support of accurate tracking in challenging scenarios.

A considerable body of research has emerged on the use of orientation selective filters in the spatiotemporal domain for the purpose of analyzing motion; a few illustrative examples follow. The general applicability of spatiotemporally oriented representations to motion perception was described some time ago, e.g., [24]. An early realized application of such ideas was to the recovery of optical flow, e.g., [25]. More recently, it has been suggested that the distribution of energy across spatiotemporal orientations is indicative of primitive patterns in visual space-time (e.g., single vs. multiple motions in a region) that can be used for qualitative distinctions in video analysis [26]. With regard to visual tracking, previous work has made use of a measure of coherent motion, as derived from oriented spatiotemporal filters, to weight image derived data in a gradient-based template tracker [27]. Also of interest is previous work that has relied on oriented, bandpass filtering purely in the spatial domain to define features for tracking, e.g., [28]. Significantly, 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 [29]. Mean shift trackers have attained much interest in recent years due to the fact that they are effective, even in the presence of clutter, partial occlusion and target deformations. Further, our proposed oriented energy features readily map onto this paradigm. (Although they also are applicable to alternative paradigms, e.g., those that preserve within target spatial relationships, as the oriented energies are calculated locally.) It appears that the mean shift algorithm was first applied to the problem of tracking by Comaniciu et al. [30]. Subsequently, various

extensions and improvements to mean shift tracking have been documented [31, 18, 32, 33, 34].

In light of previous research, the main contributions of the current approach are as follows. First, 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. Second, oriented energy features are instantiated with respect to a particular tracking mechanism, the mean shift estimator. Oriented energies map naturally onto this tracker; although, they have the potential to be applicable to a wide range of tracking paradigms. Third, the performance of the resulting system is documented both qualitatively and quantitatively. Of notable importance is the fact that our algorithm outperforms a color-based mean shift implementation in three situations that are common to real-world video sequences: substantial background clutter; multiple targets which have similar color characteristics; and during changes of illumination.

This report is organized into four major chapters. This first chapter has served to motivate the use of oriented energy-based features for tracking and to place it in the context of related research. Chapter 2 presents the details of the technical approach. Chapter 3 presents empirical evaluation. Finally, Chapter 4 serves to summarize our contributions.

Chapter 2

Technical Approach

This section details the derivation of the novel, oriented energy-based feature set that is proposed for tracking. The advantages of utilizing oriented energies in this application domain, including their robustness to illumination variation, rich description of spatiotemporal structure, and ability to track in the presence of clutter, will be further explained. The mean shift mode seeking algorithm is reviewed briefly because it is the particular tracking mechanism that was selected to illustrate the power of this feature set. Special attention is paid to the incorporation of the oriented energy features into the mean shift framework.

2.1 Oriented energy features

2.1.1 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 [24]. 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 an extremely 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 in the presence of 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 [35], 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 [36]. This Gaussian pyramid approach allows for efficient analysis of the space-time structure across multiple scales. Hence, a measure of local energy, e, can be computed according to

$$e\left(\mathbf{x};\theta,\sigma\right) = \left[G_2\left(\theta,\sigma\right) * I\left(\mathbf{x}\right)\right]^2 + \left[H_2\left(\theta,\sigma\right) * I\left(\mathbf{x}\right)\right]^2, \tag{2.1}$$

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 different orientations irrespective of local contrast, $e\left(\mathbf{x};\theta,\sigma\right)$ is normalized as

$$\hat{e}\left(\mathbf{x};\theta,\sigma\right) = \frac{e\left(\mathbf{x};\theta,\sigma\right)}{\sum_{\tilde{\sigma}}\sum_{\tilde{\theta}} e\left(\mathbf{x};\tilde{\theta},\tilde{\sigma}\right) + \epsilon}, \tag{2.2}$$

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 notational convention is to superscript variables of summation with $\tilde{\cdot}$.)

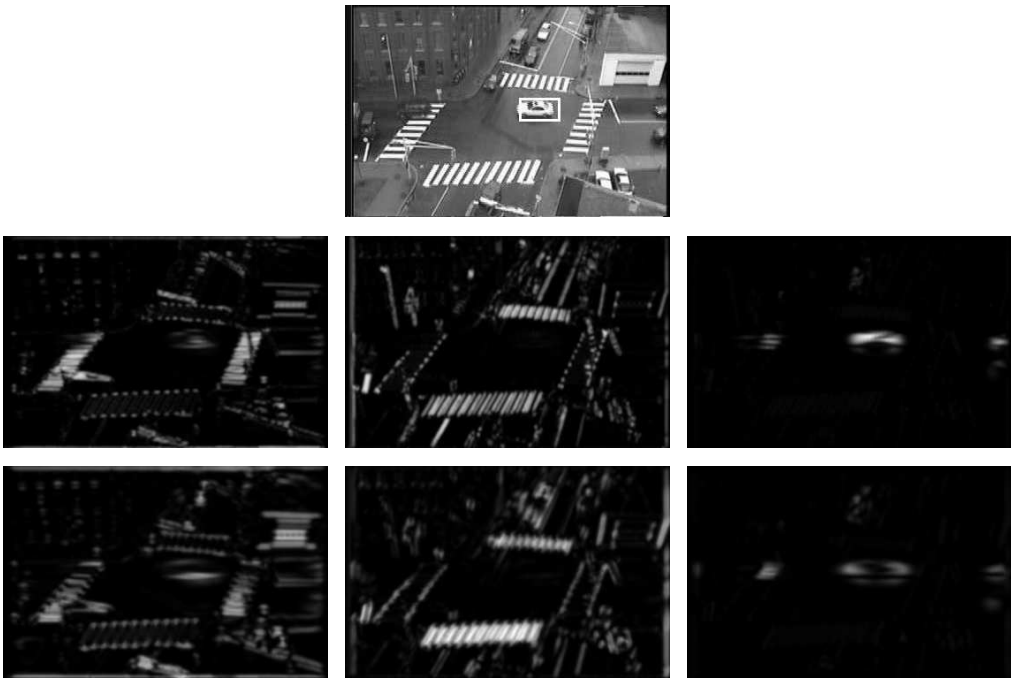

Figure 2.1: Frame 29 of the MERL traffic video sequence with select corresponding energy channels. Finer and coarser scales are shown in rows two and three, resp. From left to right, the energy channels roughly correspond to horizontal structure, vertical structure, and leftward motion.

For illustrative purposes, Fig. 2.1 displays a subset of the energies that are computed for a single frame of a MERL traffic sequence [37]. 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. 2.1, the energies computed at coarser scales are diffuse due to the downsam-

pling/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 Eqn. (2.2) according to

$$\hat{E}(\mathbf{x};\theta,\sigma) = \hat{e}(\mathbf{x};\theta,\sigma)\, b(\mathbf{x};\theta) \tag{2.3}$$

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

$$b(\mathbf{x};\theta) = \sum_{\tilde{\sigma}} \hat{e}(\mathbf{x};\theta,\tilde{\sigma}) > T_\theta. \tag{2.4}$$

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 Eqns. (2.1) and (2.2), captures local spatiotemporal structure at a particular orientation and scale with a degree of robustness to scene illumination: By virtue of the bandpass filtering, (2.1), invariance will be had to changes that are manifest in the image as additive offsets to image brightness; by virtue of the normalization, (2.2), 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 [38].

2.1.2 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 aggre-

gated over target support to provide region-based descriptors (e.g., in conjunction with mean shift tracking). Here, we pursue the second option and explore 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. The target histogram is constructed using the energies of Eqn. (2.3). Each histogram bin corresponds to the weighted energy content of the target at a particular scale and orientation. Hence, the entire histogram displays the weighted energy of the target across all scales and orientations. The energy histograms are created in a different fashion than the color histograms seen in many mean shift algorithms. For example, in [30], each pixel on the target contributes to just a single histogram bin, depending on its color. When computing our energy-based histograms, each target pixel affects every bin in the histogram. Specifically, in our tracker the template histogram which defines the target in the first frame of the video sequence is given by

$$\hat{q}_u = C \sum_{i=1}^{n} k\left(\|\mathbf{x}_i^*\|^2\right) \hat{E}\left(\mathbf{x}_i^*; \phi_u\right) \tag{2.5}$$

where k is the profile of the tracking kernel, C is a normalization constant to ensure the histogram sums to unity, $x_i^* = (x^*, y^*)$ 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. (Under our notational convention, when referring to kernels, uppercase indicates the kernel itself while lowercase refers to the kernel profile.)

When tracking a target between frames, it may be necessary to evaluate several target candidates before a final, optimal target position is found for the current frame. The histograms for the target candidates are evaluated using

$$\hat{p}_u\left(\mathbf{y}\right) = C_h \sum_{i=1}^{n_h} k\left(\left\|\frac{\mathbf{y} - \mathbf{x}_i^*}{h}\right\|^2\right) \hat{E}\left(\mathbf{x}_i^*; \phi_u\right) \tag{2.6}$$

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. The kernel bandwidth allows for scale changes of the target throughout the video sequence.

A sample energy histogram for the target region shown in Fig. 2.1 (represented by the white box) is shown in Fig. 2.2. The bin corresponding most

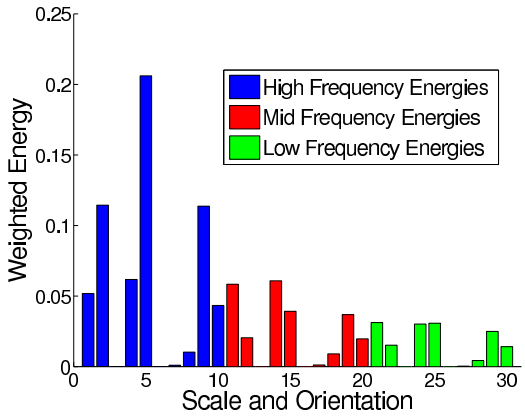

Figure 2.2: Oriented energy histogram for the target region in Fig. 2.1.

closely to leftward motion at 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

2.2.1 Target position estimation

Under the mean shift framework, tracking an object essentially involves locating the candidate position in the current frame that produces the histogram that is most similar to the template histogram. Thus, a measure of similarity between two histograms is required. We utilize the Bhattacharyya coefficient for histogram comparisons. The sample estimate of the Bhattacharyya coef-

ficient can be computed according to

$$\rho\left[\hat{\mathbf{p}}\left(\mathbf{y}\right),\hat{\mathbf{q}}\right]=\sum_{u=1}^{m}\sqrt{\hat{p}_u\left(\mathbf{y}\right)\hat{q}_u}. \tag{2.7}$$

where $\hat{\mathbf{p}}\left(\mathbf{y}\right)$ and $\hat{\mathbf{q}}$ histogram bins of cardinality m. Due to the definition of the Bhattacharyya coefficient, in order to minimize the distance between two histograms, Eqn. (2.7) must be maximized with respect to target position $\mathbf{y}$.

The Bhattacharyya coefficient can be maximized via mean shift iterations [39, 29]. The specific mean shift vector that can be used to perform the desired maximization is

$$\hat{\mathbf{y}}_1=\left[\frac{\sum_{i=1}^{n_h}\mathbf{x}_i^* w_i g\left(\left|\left|\frac{\hat{\mathbf{y}}_0-\mathbf{x}_i^*}{h}\right|\right|^2\right)}{\sum_{i=1}^{n_h} w_i g\left(\left|\left|\frac{\hat{\mathbf{y}}_0-\mathbf{x}_i^*}{h}\right|\right|^2\right)}\right] \tag{2.8}$$

where

$$w_i=\sum_{u=1}^{m}\hat{E}\left(\mathbf{x}_i^*;\phi_u\right)\sqrt{\frac{\hat{q}_u}{\hat{p}_u\left(\hat{\mathbf{y}}_0\right)}}, \tag{2.9}$$

$g\left(x\right)=k'\left(x\right)$ is the derivative with respect to x of tracking kernel profile, k, and $\hat{\mathbf{y}}_0$ is the current position of the target. Here, the Epanechnikov kernel has been shown to be effective [34, 30, 32] and appears to be the most commonly used kernel in application of mean shift to computer vision.

Thus, the position of the target in the current frame is estimated as follows. Starting from the target's position in the previous frame, iterations are performed whereby 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.

2.2.2 Template and scale updates

When tracking an object through a long video sequence, it is common that its characteristics will change. For example, with color features, if there is a change in illumination, it may be desired to update the target template to accurately reflect the target's appearance under the new lighting conditions. Similarly, when using oriented energy-based features, the template may need to be updated if the target's energy distribution changes. Such changes may

be caused by alterations in velocity or rotations. To combat the changes a target may incur over time, our tracker incorporates a simple template update mechanism defined as

$$\hat{\mathbf{q}}^{i+1} = \alpha\pi\hat{\mathbf{q}}^{i} + (1-\alpha)(1-\pi)\,\hat{\mathbf{p}}(\mathbf{y}_i), \tag{2.10}$$

where α is a weighting factor to control the speed of template updates, $\hat{\mathbf{q}}^{i}$ is the template at frame i, and $\pi = \rho[\hat{\mathbf{p}}(\mathbf{y}_i), \hat{\mathbf{q}}^{i}]$ 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. Also, following each application of (2.10), the resulting template distribution 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. As the difference between the two histograms becomes greater, the template will be modified so that it is more similar to the optimal candidate.

In addition to experiencing alterations to their energy distributions, the size of a target may change during a video sequence as well. Although there are more effective methods of dealing with changes in object scale in the mean shift framework [31, 33], in the current implementation we employ a simple approach, similar to that taken in [30]. 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 considered in this work. 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

$$h_{new} = \gamma h_{opt} + (1-\gamma)\,h_{prev}. \tag{2.11}$$

Empirically, we set $\gamma = 0.15$

Chapter 3

Empirical Evaluation

The proposed oriented energy features have been incorporated into a software implementation of a mean shift tracker. The performance of the resulting system 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. Apart from the fact that the different histograms were used, the two trackers were identical. The color-based tracker was implemented in a similar manner to [30], 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 different scales with 10 different spatiotemporal orientations per scale. Accordingly, the energy-based histograms contained 30 bins. Note the compactness of the oriented energy representation when compared against the color-based features. For the oriented energy feature set, 10 orientations were selected as they span the space of 3D orientation for the highest order filters that we use (H_2) [35]; 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 [40]. For all results presented in this paper, an Epanechnikov kernel, K, was used. The thresholds for Eqn. 2.4 were empirically set to be $2.75\times$ the mean energy for each orientation channel. In the subsequent experiments, both the color and the energy-based trackers were hand-initialized with identical target regions in the initial frame of each video.

The first video sequence from the test set illustrates the effectiveness of oriented energy-based features in dealing with illumination changes. In particular, an individual starts walking in a poorly lit area; then, he travels

into and out of the bright region as he walks across the room. Results for both color and energy-based trackers are shown in Fig. 3.1. The room, the background, and the individual's clothing are all very dark and thus, they appear very similar under a color-based representation. As Fig. 3.1 illustrates, the color-based mean shift tracker completely looses track of the target after only a few frames. While it might be possible for color-based tracking representations to evolve with illumination changes [16], relevant techniques typically have not been incorporated into general-purpose mean shift trackers. Some approaches, such as the scaling of the RGB space [15] or using the HS components of the HSV color space [15] have been considered, but primarily for the application of tracking skin-colored regions. To ensure unbiased evaluation, we also ran the color-based mean shift tracker using histograms created using normalized RG-space [29] and found that it still lost track after approximately ten frames. In contrast to the poor performance of the color-based tracker, our proposed feature set enabled the system to follow the walker extremely well. The performance of the oriented energy-based mean shift tracker qualitatively appeared to be relatively unaffected by the changes in illumination. This robustness arises from the normalization performed in Equation (2.2).

Figure 3.1: Video sequence ($x \times y \times t = 360 \times 240 \times 60$) of a man walking through shadows. Results from the color-based tracker are on the top; images when oriented energy histograms were used are on the bottom. From left to right, frames 4, 18, 31, and 55 are shown. Tracked regions are highlighted with white boxes.

Experiments were also completed on video sequences where multiple individuals with similar colored clothing walk in the vicinity of one another.

Fig. 3.2 displays a video sequence where two individuals are walking in opposite directions in a room. The clothing of both individuals is very similar in color, but the patterns on their shirts' are different. The results of the color-based algorithm are shown in Fig. 3.2 when the individual who starts walking on the right side of the video is being tracked. Just prior to the occlusion event, the color-based tracker was successfully following the correct individual. However, during the occlusion of the true target, the color-based tracker becomes distracted by the other walker. This confusion occurs because the two individuals in the video are almost identical under a color-based representation. The strength of the energy-based features becomes apparent when one considers the second set of results in the bottom row of Fig. 3.2. The oriented energy features representing the target span the spatiotemporal domain; thus both the target's appearance and its motion are described by the proposed energy-based histograms. In the video of Fig. 3.2, it can be seen that despite the full occlusion that occurs for several frames, the tracker using energy features is capable of following the true target walker until the end of the video. The different texture patterns and velocities of the walkers were sufficient cues for the oriented energy-based tracker to achieve success.

Figure 3.2: Video sequence ($x \times y \times t = 360 \times 240 \times 50$) of people walking through a room with similar colored clothing. Results from the color-based tracker are on the top; images when oriented energy histograms were used are on the bottom. From left to right, frames 6, 18, 32, and 50 are shown. Tracked regions are highlighted with white boxes.

In addition to comparing favorably against a color-based mean shift tracker in the videos of Fig. 3.1 and 3.2, our proposed feature set has shown to be ca-

pable of tracking objects when applied to a wide array of additional videos. Fig. 3.3 shows a real-life, grayscale video sequence of a traffic intersection that was obtained from MERL [37]. This video sequence exhibits clutter in the form of traffic lights, lampposts, and several moving vehicles at the intersection. 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 decreased level of performance occurs because as the car traverses the crosswalk, the lack of contrast (essentially uniform white on white) between foreground and background yields little energy for the involved portions of the car. Accordingly, the tracker searches elsewhere for a better match. Nevertheless, the tracker never loses the target; indeed, the frames shown are representative of the worst case performance in this video. It also should be noted that a color-based tracker would be challenged by the presence of the crosswalk since it has a similar color distribution to the car.

Figure 3.3: 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.

Our feature set was also successfully used when tracking people and vehicles in videos obtained from the PETS2001 dataset [41]. Fig. 3.4 shows an example 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. It also should be noted that the cars in the scene add clutter to the background and that some of these cars have similar color distributions to that of the cyclist. 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, during the occlusion (results not shown).

Figure 3.4: 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.

Results from another illustrative video sequence are shown in Fig. 3.5. Here, the individual is walking erratically, making sudden changes in direction and moving at a wide variety of speeds. Since the oriented 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 update, (2.10), incorporates changes to adapt the model for further tracking.

Figure 3.5: 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.

Fig. 3.6 displays the tracking performance using the proposed features for a final sequence. This video is representative of the footage that one might obtain from overhead surveillance cameras in public areas. The oriented energy-based tracker is capable of following the target of interest even though there are multiple walking individuals that have a similar appearance. The complexity of the sequence is further increased because the target has little texture on his clothing. The shadows that are cast from trees outside the field of view of the camera present yet another challenge when tracking.

Finally, reflective and semi-transparent effects are visible in the video, particularly on the right side of the frame, because it was recorded through a window. When tracking is performed using the oriented energy feature set, the target is not lost, even during the partial occlusion that occurs from approximately frame 16 - 38. Indeed, the tracker does lag behind the target for a few frames immediately following the occlusion as it decides which person provides the best match to the template. However, frame 39 is representative of its worst-case performance in this video. In comparison, our color-based implementation was only able to follow the true target for approximately 30 frames before it locked on to another walker in the scene (results not shown).

Figure 3.6: 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.

Quantitative performance analysis was performed for the video sequences that are publicly available — MERL and PETS2001. Specifically, Fig. 3.7

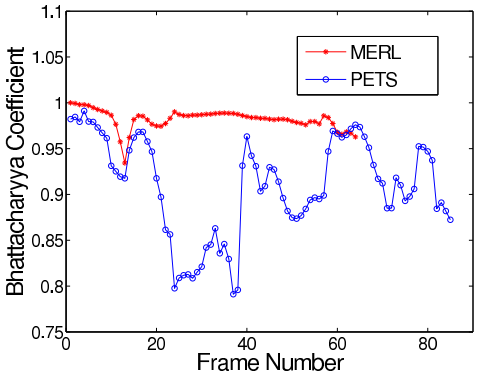

Figure 3.7: Bhattacharyya coefficients over the entire video sequence for the MERL and PETS2001 videos.

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. 3.7 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.

Chapter 4

Summary

Spatiotemporal oriented energy features provide a rich, yet compact representation of a targets 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. Significantly, oriented energies also could provide the representational substrate for other tracking mechanisms, e.g., spatial template-based methods by providing pointwise measurements of template structure for matching/warping across time. In the current implementation, the approach shares similarities with other mean shift-based trackers. The primary and significant difference in our instantiation is the use of an oriented energy-based histogram representation, rather than more standard descriptors, e.g., color-based. Within this framework, empirical results show that oriented energy features support accurate tracking, as our empirical evaluations over a wide range of video sequences have shown. In our experiments, 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 video sequences that displayed substantial background clutter, targets which contained similar colors to other objects in the scene, and changes in illumination. These difficulties are significant and are often present in real-world surveillance video sequences. Tracking with the use of oriented energy features was shown to be robust to the aforementioned challenging conditions.

Bibliography

[1] Yilmaz, A., Javed, O., Shah, M.: Object tracking: A survey. Comp. Surv. **38**(4) (2006) 1–45

[2] Lucas, B., Kanade, T.: An iterative image registration technique with application to stereo vision. In: DARPA IUW. (1981) 121–130

[3] Burt, P., Yen, C., Xu, X.: Local correlation measures for motion analysis: A comparative analysis. In: PRIP. (1982) 269–274

[4] Anandan, P.: A computational framework and an algorithm for the measurement of visual motion. IJCV **2**(3) (1989) 283–310

[5] Shi, J., Tomasi, C.: Good features to track. CVPR **1** (1994) 593–600

[6] Dreschler, L., Nagel, H.: Volumetric model and 3d trajectory of a moving car derived from monocular TV frame sequences of a street scene. In: IJCAI. (1981) 692–697

[7] Sethi, I., Jain, R.: Finding trajectories of feature points in monocular images. PAMI **9**(1) (1987) 56–73

[8] Deriche, R., Faugeras, O.: Tracking line segments. IVC **8**(4) (1991) 261–270

[9] Rangarajan, K., Shah, M.: Establishing motion correspondence. CVGIP **54**(1) (1991) 56–73

[10] Huttenlocher, D., Noh, J., Rucklidge, W.: Tracking nonrigid objects in complex scenes. In: ICCV. (1993) 93–101

[11] Terzopoulos, D., Szeliski, R.: Tracking with kalman snakes. In Blake, A., Yuille, A., eds.: Active Vision. MIT Press (1992) 553–556

[12] Isard, M., Blake, A.: Contour tracking by stochastic propagation of conditional density. ECCV **1** (1996) 343–354

[13] Haritaoglu, L., Harwood, D., Davis, L.: W4: Real-time surveillance of people and their activities. PAMI **22**(8) (2000) 809–830

[14] Fieguth, P., Terzopoulos, D.: Color-based tracking of heads and other mobile objects at video frame rates. In: CVPR. (1997) 21–27

[15] Birchfield, S.: Elliptic head tracking with intensity gradients and color histograms. CVPR **1** (1998) 232–237

[16] Sigal, L., Sclaroff, S., Athitsos, V.: Estimation and prediction of evolving color distributions for skin segmentation under varying illumination. CVPR **2** (2000) 152–159

[17] Wu, Y., Huang, T.: A co-inference approach to robust visual tracking. In: ICCV. (2001) 26–33

[18] Elgammal, A., Duraiswami, R., Davis, L.: Probabilistic tracking in joint feature-spatial spaces. CVPR **1** (2003) 781–788

[19] Bolgomolov, Y., Dror, G., Lapchev, S., Rivlin, E., Rudzsky, M.: Classification of moving targets based on motion and appearance. In: BMVC. (2003) 142–149

[20] Cremers, D., Schnorr, C.: Statistical shape knowledge in variational motion segmentation. IVC **21**(1) (2003) 77–86

[21] Sato, K., Aggarwal, J.: Temporal spatio-velocity transformation and its application to tracking and interaction. CVIU **96**(2) (2004) 100–128

[22] Avidan, S.: Support vector tracking. In: CVPR. (2001) 184–191

[23] Okuma, K., Taleghani, A., Freitas, N.D., Little, J., Lowe, D.: A boosted particle filter: multitarget detection and tracking. ECCV **1** (2004) 28–39

[24] Adelson, E., Bergen, J.: Spatiotemporal energy models for the perception of motion. JOSA A **2**(2) (1985) 284–299

[25] Heeger, D.: Optical flow from spatiotemporal filters. IJCV **1**(4) (1988) 297–302

[26] Wildes, R., Bergen, J.: Qualitative spatiotemporal analysis using an oriented energy representation. ECCV **2** (2000) 784–796

[27] Enzweiler, M., Wildes, R., Herpers, R.: Unified target detection and tracking using motion coherence. Wrkshp. Motion & Video Comp. **2** (2005) 66–71

[28] Jepson, A., Fleet, D., El-Maraghi, T.: Robust online appearance models for visual tracking. IEEE PAMI **25**(10) (2003) 1296–1311

[29] Comaniciu, D., Ramesh, V., Meer, P.: Kernel-based object tracking. IEEE PAMI **25**(5) (2003) 564–575

[30] Comaniciu, D., Ramesh, V., Meer, P.: Real-time tracking of non-rigid objects using mean shift. CVPR **2** (2000) 142–149

[31] Collins, R.: Mean-shift blob tracking through scale space. CVPR **2** (2003) 234–240

[32] Hager, G., Dewan, M., Stewart, C.: Multiple kernel tracking with SSD. CVPR **1** (2004) 790–797

[33] Zivkovic, Z., Krose, B.: An EM-like algorithm for color-histogram tracking. CVPR **1** (2004) 798–803

[34] Birchfield, S., Rangarajan, S.: Spatiograms versus histograms for region-based tracking. CVPR **2** (2005) 1158–1163

[35] Freeman, W., Adelson, E.: The design and use of steerable filters. IEEE PAMI **13**(9) (1991) 891–906

[36] Burt, P., Adelson, E.: The Laplacian pyramid as a compact image code. IEEE TC **31**(4) (1983) 532–540

[37] Brand, M., Kettnaker, V.: Discovery and segmentation of activities in video. IEEE PAMI **22**(8) (2000) 844–851

[38] Derpanis, K., Gryn, J.: Three-dimensional nth derivative of Gaussian separable steerable filters. ICIP **3** (2005) 553–556

[39] Fukunaga, K., Hostetler, L.: The estimation of the gradient of a density function, with applications in pattern recognition. IEEE IT **21**(1) (1975) 32–40

[40] Pearce, P., Pearce, S.: Polyhedra Primer. van Nostrand Reinhold Company, New York, New York (1978)

[41] PETS. http://peipa.essex.ac.uk/ipa/pix/pets/ (2006)