Action Spotting and Recognition Based on a Spatiotemporal Orientation Analysis

Konstantinos G. Derpanis, *Member*, *IEEE*, Mikhail Sizintsev, *Member*, *IEEE*, Kevin J. Cannons, and Richard P. Wildes, *Member*, *IEEE*

Abstract—This paper provides a unified framework for the interrelated topics of *action spotting*, the spatiotemporal detection and localization of human actions in video, and *action recognition*, the classification of a given video into one of several predefined categories. A novel compact local descriptor of video dynamics in the context of action spotting and recognition is introduced based on visual spacetime oriented energy measurements. This descriptor is efficiently computed directly from raw image intensity data and thereby forgoes the problems typically associated with flow-based features. Importantly, the descriptor allows for the comparison of the underlying dynamics of two spacetime video segments irrespective of spatial appearance, such as differences induced by clothing, and with robustness to clutter. An associated similarity measure is introduced that admits efficient exhaustive search for an action template, derived from a single exemplar video, across candidate video sequences. The general approach presented for action spotting and recognition is amenable to efficient implementation, which is deemed critical for many important applications. For action spotting, details of a real-time GPU-based instantiation of the proposed approach are provided. Empirical evaluation of both action spotting and action recognition on challenging datasets suggests the efficacy of the proposed approach, with state-of-the-art performance documented on standard datasets.

Index Terms—Action spotting, action recognition, action representation, human motion, visual spacetime, spatiotemporal orientation, template matching, real-time implementations

1 Introduction

1.1 Motivation

This paper addresses the interrelated topics of *detecting and localizing* spacetime patterns in a video and *recognizing* spacetime patterns. Specifically, patterns of current concern are those induced by human actions. Here, "action" refers to a simple dynamic pattern executed by an actor over a short duration of time (e.g., walking and hand waving). In contrast, activities can be considered as compositions of actions, sequentially, in parallel, or both. Potential applications of the presented research include video indexing and browsing, surveillance, visually guided interfaces, and tracking initialization.

Joint detection and localization of actions is herein referred to as "action spotting" (cf. word spotting in speech recognition). Action spotting seeks to detect and spatiotemporally localize an action, represented by a small video clip (i.e., the query), within a larger video that may contain a large corpus of unknown actions. In the present work, action spotting is achieved by a *single* query video that defines the action template, rather than a training set of (positive and negative) exemplars. In contrast, action recognition assigns a video segment to an action category taken from a set of predefined actions. For example, evaluation of recognition performance on the publicly available KTH [1], UCF Sports [2], and Hollywood2 [3] action benchmarks assigns a query video to one of 6, 10, and 12 categories, respectively. Typically, action spotting and recognition have been considered in a disjoint manner, whereas the presented approach handles action spotting and recognition in a common framework.

A key challenge in both action spotting and recognition arises from the fact that the same action-related pattern dynamics can yield very different image intensities due to spatial appearance differences, as with changes in clothing. Another challenge arises in natural imaging conditions where scene clutter requires the ability to distinguish relevant pattern information from distractions. Clutter can be of two types: 1) Background clutter arises when actions are depicted in front of complicated, possibly dynamic, backdrops and 2) foreground clutter arises when actions are depicted with distractions superimposed, as with dynamic lighting, pseudotransparency (e.g., walking behind a chain-link fence), temporal aliasing, and weather effects (e.g., rain and snow). It is proposed that the choice of representation is key to meeting these challenges: A representation that is invariant to purely spatial pattern allows actions to be recognized independent of actor appearance; a representation that supports fine delineations of spacetime structure makes it possible to tease action information from clutter. Also, for real-world applications such as video retrieval from the web, computational efficiency is a further requirement.

For the present purposes, local spatiotemporal orientation is of fundamental descriptive power as it captures the first-order correlation structure of the data irrespective of its origin (i.e., irrespective of the underlying visual phenomena), even while distinguishing a wide range of

• *The authors are with the Department of Computer Science and Engineering, York University, CSB 1003, 4700 Keele St., Toronto, Ontario M3J 1P3, Canada.*
E-mail: {kosta, sizints, kcannons, wildes}@cse.yorku.ca.

Manuscript received 16 Dec. 2010; revised 6 Nov. 2011; accepted 19 May 2012; published online 20 June 2012.
Recommended for acceptance by I. Reid.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2010-12-0964.
Digital Object Identifier no. DOI: 10.1109/TPAMI.2012.141.

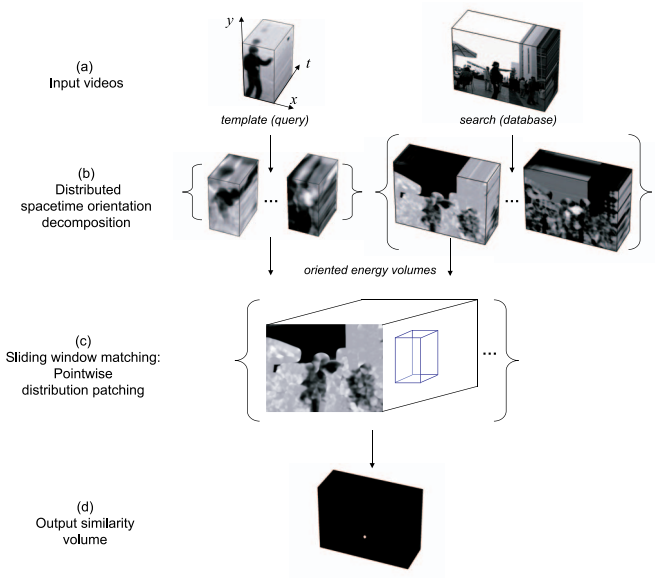


Fig. 1. Overview of approach to action spotting. (a) A template (query) containing the isolated action of interest and search (database) video serve as input; the template and search videos in the figure depict a boxing action taken from the KTH [1] and MSR [10] action datasets, respectively. (b) Application of spacetime oriented energy filters decomposes the input videos into a distributed representation according to 3D, (x, y, t), spatiotemporal orientation. (c) In a sliding window manner, the distribution of oriented energies of the template is compared to the search distribution at corresponding positions to yield the similarity volume given in (d). Finally, significant local maxima in the

image dynamics (e.g., single motion, multiple superimposed motions [4], and temporal flicker). Correspondingly, visual spacetime will be represented according to its local 3D, (x, y, t), orientation structure: Each point of spacetime will be associated with a distribution of measurements indicating the relative presence of a particular set of spatiotemporal orientations. Comparisons in searching are made between these distributions.

Fig. 1 provides an overview of the proposed action spotting approach. For action recognition, an *uncropped* video constitutes the query video and a set of labeled video snippets containing spatiotemporally localized actions form the database. The query video is compared to each action in the database and the label of the action with the global maximum similarity value is returned as the category (cf., [5], [6], [7], [8]). In addition, the proposed approach provides spatiotemporal localization information; note that the issue of localization has generally been ignored in action recognition related work. A preliminary description of this work has appeared previously [9].

1.2 Related Work

A wealth of work has considered the analysis of human actions from visual data, e.g., [11], [12]. One manner of organizing this literature is in terms of the underlying representation of actions. A brief corresponding survey of representative approaches follows.

Tracking-based methods begin by tracking body parts, joints, or both and classify actions based on features extracted from the motion trajectories, e.g., [13], [14], [15], [16]. General impediments to fully automated operation include tracker initialization and robustness. Consequently, much of this work has been realized with some degree of human intervention.

Other methods have classified actions based on features extracted from 3D spacetime body shapes as represented by contours or silhouettes, with the motivation that such representations are robust to spatial appearance details [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27]. This class of approach relies on figure-ground segmentation across spacetime, with the drawback that robust segmentation remains elusive in uncontrolled settings. Further, silhouettes do not provide information on the human body limbs when they are in front of the body (i.e., inside the silhouette) and thus yield ambiguous information.

Recently, spacetime interest points [28], [29], [30], [31], [32], [33] have emerged as a popular means for action recognition [1], [34], [35], [36], [37], [38], [39], [10]. Interest points typically are taken as spacetime loci that exhibit variation along all spatiotemporal dimensions and provide locations for extracting descriptors capturing dynamics and spatial appearance. Commonly, these descriptors have been combined to define a global video descriptor (e.g., bag of visual words). Sparsity is appealing as it yields significant reduction in computational effort; however, interest point detectors often fire erratically on shadows and highlights [40], [41], as well as along object occluding boundaries, and may be overwhelmed by the background-related interest points in highly dynamic situations, which casts doubt on their applicability to cluttered natural imagery. Additionally, for actions substantially comprised of smooth motion, important information is ignored in favor of a small number of possibly nondiscriminative interest points. To ameliorate some of these issues, various recent works have relied on interest point schemes that recover a denser set of points across the video (e.g., [42]) or forgo the use of interest points all together and instead compute the descriptor at each image point [43], [44], [45], [46].

Most closely related to the approach proposed in the present paper are others that have considered dense templates of image-based measurements to represent actions (e.g., intensity image, optical flow, spatiotemporal gradients, and other filter responses selective to both spatial and temporal orientation). Typically, these measurements are matched to a video of interest in a sliding window formulation. The chief advantages of this framework include avoiding problematic preprocessing of the input video such as localization, tracking, and segmentation; however, such approaches can be computationally intensive. Further limitations are tied to the particulars of the image measurement used to define the template.

Approaches have avoided preprocessing the raw input imagery and used it directly as the initial representation, e.g., [47]. This tack places the burden on the learning process to abstract the action-related features. Alternatively, more abstracted features (e.g., optical flow, gradients, etc.) can serve as the basis for matching. In general, optical flow-based methods, e.g., [48], [40], [49], [2], [50], [26], [51], suffer as dense flow estimates are unreliable where their local single flow assumption does not hold (e.g., along occluding boundaries and in the presence of foreground clutter).

Work using spatiotemporal gradients has encapsulated the measurements in the gradient structure tensor [20], [52], [53], [8]. This tack yields a compact way to characterize visual spacetime locally, with template video matches computed via dimensionality comparisons. However, the compactness also limits its descriptive power: Areas containing two or more orientations in a region are not readily discriminated, as their dimensionality will be the same; further, the presence of foreground clutter in a video of interest will contaminate dimensionality measurements to yield match failures. Finally, methods based on filter responses selective for both spatial and temporal orientation, e.g., [54], [55], [56], [7], suffer from their inability to generalize across differences in spatial appearance of the same action, such as differences in clothing.

Many of the extant approaches have focused on high detection accuracy while giving little attention to computational efficiency issues. As in the current paper, several recent works have specifically addressed the computational efficiency aspects of action recognition [53], [37], [23], [25], [26], [43], [42], [10].

Also similar to the current work, several methods have focused on recognizing actions based on a *single query*, where a query video is encapsulated in a sliding-window and compared against an annotated database of videos [20], [52], [47], [7], [8]. These data-driven methods may prove particularly useful for video retrieval tasks. For example, in situations where a user provides a single video snippet of an action, an applicable automated approach must be able to return the most similar instances in a video database without the luxury of additional positive or negative examples (cf. Google's "Search by Image" service [57]).

The features used in the present work derive from spatiotemporal oriented filtering that captures dynamic aspects of visual spacetime with robustness to purely spatial appearance. Previous work in optical flow estimation has made use of spatiotemporal filtering that discounts spatial appearance in a different fashion than employed in the current paper, as it appealed to a nonlinear optimization procedure [58]. More closely related is previous work that used the same filtering techniques employed in the present paper, with two significant differences. First, these efforts applied the filtering to very different research domains of video segmentation [59] and dynamic texture recognition [60]. Second, the previous work aggregated the filter responses over relatively large regions of support to yield a single distribution of measurements to represent a region of interest, whereas in the present work each point in a dense action template is associated with its own distribution to maintain the spatiotemporal organization of the action. Significantly, it appears that the present contribution is the first to apply and demonstrate the usefulness of the proposed spatiotemporal filtering approach to action analysis in any fashion.

1.3 Contributions

In the light of previous work, the major contributions of the present paper are as follows:

1. A novel compact local oriented energy feature set is developed for action spotting and recognition. This representation supports fine delineations of visual spacetime structure to capture the rich underlying dynamics of an action.
2. An associated computationally efficient similarity measure and search method are proposed that leverage the structure of the representation. The approach does not require preprocessing in the form of actor localization, tracking, motion estimation, or figure-ground segmentation.
3. The approach can accommodate variable appearance of the same action, rapid dynamics, multiple actions in the field-of-view, cluttered backgrounds and is resilient to the addition of distracting foreground clutter. While others have dealt with background clutter, it appears that the present work is the first to address directly the foreground clutter challenge.
4. A real-time implementation of the action spotting approach is documented.
5. The proposed approaches to action spotting and recognition are evaluated on a wide variety of challenging videos, with state-of-the-art performance documented on standard datasets.

2 Technical Approach

In visual spacetime, the local 3D, (x, y, t), orientation structure of a pattern captures significant, meaningful aspects of its dynamics. For action spotting and recognition, single motion at a point, e.g., motion of an isolated body part, is captured as orientation along a particular spacetime direction. Significantly, more complicated scenarios still give rise to well-defined spacetime orientation distributions: Occlusions and multiple motions (e.g., as limbs cross or foreground clutter intrudes) correspond to multiple orientations; high velocity and temporal flicker (e.g., as encountered during rapid action executions) correspond to orientations that become orthogonal to the temporal axis. Further, appropriate definition of local spatiotemporal oriented energy measurements can yield invariance to purely spatial pattern characteristics and support action analysis as an actor changes spatial appearance. Based on these observations, the action spotting and recognition approaches developed make use of spatiotemporal orientation measurements as local features that are combined into spacetime templates that maintain their relative spacetime positions.

In this work, it is assumed that the camera is stationary in order for the proposed spacetime measurements to capture the dynamics of the action rather than the camera movement. Empirically, the proposed features have been found to be robust to small amounts of camera movement, such as camera jitter from a handheld video camcorder; however, they are not invariant to large camera movements. To accommodate large camera movements, a camera stabilization procedure may be introduced as a preprocessing step, e.g., [37].

2.1 Features: Spatiotemporal Orientation

The desired spatiotemporal orientation decomposition is realized using broadly tuned 3D Gaussian third derivative filters, $G_{3_{\hat{\theta}}}(\mathbf{x})$, with the unit vector $\hat{\theta}$ capturing the 3D direction of the filter symmetry axis and $\mathbf{x} = (x, y, t)$ the

spacetime position. The responses of the image data to these filters are pointwise rectified (squared) and integrated (summed) over a spacetime neighborhood, Ω, to yield the following locally aggregated pointwise energy measurements:

$$E_{\hat{\theta}}(\mathbf{x}) = \sum_{\mathbf{x}\in\Omega} (G_{3_{\hat{\theta}}} * I)^2, \tag{1}$$

where $I \equiv I(\mathbf{x})$ denotes the input imagery and $*$ convolution. Notice that while the employed Gaussian derivative filters are phase-sensitive, summation over the support region ameliorates this sensitivity to yield a measurement of signal energy at orientation $\hat{\theta}$. More specifically, this follows from Rayleigh's-Parseval's theorem [61] that specifies the phase-independent signal energy in the frequency passband of the Gaussian derivative:

$$E_{\hat{\theta}}(\mathbf{x}) \propto \sum_{\omega_x,\omega_y,\omega_t} |\mathcal{F}\{G_{3_{\hat{\theta}}} * I\}(\omega_x, \omega_y, \omega_t)|^2, \tag{2}$$

where (ω_x, ω_y) denote the spatial frequency, ω_t the temporal frequency, and $\mathcal{F}$ the Fourier transform.[1]

Each oriented energy measurement, (1), is confounded with spatial orientation. Consequently, in cases where the spatial structure varies widely about an otherwise coherent dynamic region (e.g., single motion of a surface with varying spatial texture), the responses of the ensemble of oriented energies will reflect this behavior and thereby are spatial appearance dependent, whereas a description of pure pattern dynamics is sought. Note that while in tracking applications it is vital to preserve both the spatial appearance and dynamic properties of a region of interest, in action spotting and recognition it is desirable to be invariant to appearance while being sensitive to dynamic properties. This quality is necessary so as to detect different people wearing a variety of clothing as they perform the same action. To remove this difficulty, the spatial orientation component is discounted by "marginalization" as follows.

In general, a pattern exhibiting a single spatiotemporal orientation (e.g., image velocity) manifests itself as a plane through the origin in the frequency domain [62], [63]. Correspondingly, summation across a set of x-y-t-oriented energy measurements consistent with a single plane through the origin in the frequency domain is indicative of energy along the associated spatiotemporal orientation, independent of purely spatial orientation. Since Gaussian derivative filters of order $N = 3$ are used in the oriented filtering, (1), it is appropriate to consider $N + 1 = 4$ equally spaced directions along each frequency domain plane of interest, as $N + 1$ directions are needed to span orientation in a plane with Gaussian derivative filters of order N [64]. Let each plane be parameterized by its unit normal, $\hat{\mathbf{n}}$; a set of equally spaced $N + 1$ directions within the plane is given as

$$\hat{\theta}_i = \cos\left(\frac{\pi i}{N+1}\right)\hat{\theta}_a(\hat{\mathbf{n}}) + \sin\left(\frac{\pi i}{N+1}\right)\hat{\theta}_b(\hat{\mathbf{n}}), \tag{3}$$

with $\hat{\theta}_a(\hat{\mathbf{n}}) = \hat{\mathbf{n}} \times \hat{\mathbf{e}}_x / \|\hat{\mathbf{n}} \times \hat{\mathbf{e}}_x\|$, $\hat{\theta}_b(\hat{\mathbf{n}}) = \hat{\mathbf{n}} \times \hat{\theta}_a(\hat{\mathbf{n}})$, $\hat{\mathbf{e}}_x$ the unit vector along the ω_x-axis,[2] and $0 \le i \le N$.

Now, energy along a frequency domain plane with normal $\hat{\mathbf{n}}$ and spatial orientation discounted through marginalization is given by summing across the set of measurements, $E_{\hat{\theta}_i}$, as

$$\tilde{E}_{\hat{\mathbf{n}}}(\mathbf{x}) = \sum_{i=0}^{N} E_{\hat{\theta}_i}(\mathbf{x}), \tag{4}$$

with $\hat{\theta}_i$ one of $N + 1 = 4$ specified directions, (3), and each $E_{\hat{\theta}_i}$ calculated via the energy filtering, (1).

Finally, the marginalized energy measurements, (4), are confounded by the local contrast of the signal and as a result, increase monotonically with contrast. This sensitivity makes it impossible to determine whether a high response for a particular spatiotemporal orientation is indicative of its presence or is indeed a low match that yields a high response due to significant contrast in the signal. To arrive at a purer measure of spatiotemporal orientation, the energy measures are normalized by the sum of consort planar energy responses at each point:

$$\hat{E}_{\hat{\mathbf{n}}_i}(\mathbf{x}) = \tilde{E}_{\hat{\mathbf{n}}_i}(\mathbf{x}) \Bigg/ \left(\sum_{j=1}^{M} \tilde{E}_{\hat{\mathbf{n}}_j}(\mathbf{x}) + \epsilon\right), \tag{5}$$

where M denotes the number of spatiotemporal orientations considered and ϵ is a constant introduced as a noise floor and to avoid instabilities at points where the overall energy is small. As applied to the M oriented, appearance marginalized energy measurements, (4), (5) produces a corresponding set of M normalized, marginalized oriented energy measurements. To this set an additional measurement is included that explicitly captures lack of structure (i.e., untextured regions, such as a clear sky) via the normalized ϵ:

$$\hat{E}_{\epsilon}(\mathbf{x}) = \epsilon \Bigg/ \left(\sum_{j=1}^{M} \tilde{E}_{\hat{\mathbf{n}}_j}(\mathbf{x}) + \epsilon\right), \tag{6}$$

to yield an $M + 1$ dimensional feature vector at each point in the image data. Note for loci where oriented structure is less apparent, the summation in (6) will tend to 0; hence, $\hat{E}_{\epsilon}$ approaches 1 and thereby indicates relative lack of structure. The normalized energy measurements are taken together as a distribution parameterized by spatiotemporal orientation, $\hat{\mathbf{n}}$. In practice, the set of measurements are maintained as a $(M + 1)$-D histogram.

Conceptually, (1)-(6) can be thought of as taking an image sequence and carving its (local) power spectrum into a set of planes, with each plane corresponding to a particular spatiotemporal orientation, to provide a relative indication of the presence of structure along each plane or lack thereof in the case of a uniform intensity region as captured by the normalized ϵ, (6). This orientation decomposition of input imagery is defined pointwise in spacetime. For present purposes, it is used to define spatiotemporally dense 3D, (x, y, t), action templates from an example video (with each point in the template associated with a $(M + 1)$-D orientation

1. Strictly, Rayleigh's theorem is stated with infinite frequency domain support on summation.

2. Depending on the spatiotemporal orientation sought, $\hat{\mathbf{e}}_x$ can be replaced with another axis to avoid an undefined vector.

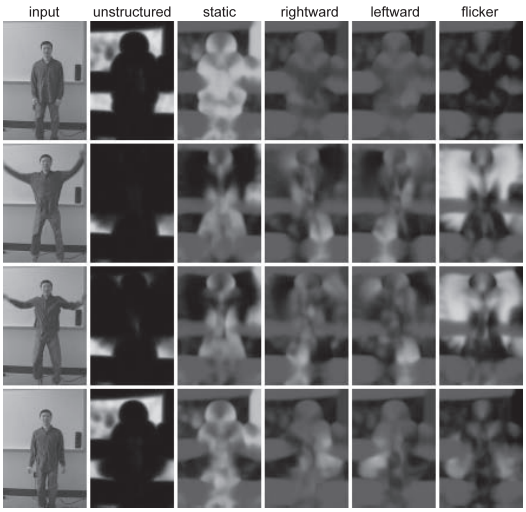


Fig. 2. Input frames of a jumping jack sequence with select corresponding energy channels from left-to-right, respectively; high intensities in the energy channels denote high orientation response. The energy channels are tuned to regions devoid of structure (unstructured), no motion but textured (static), rightward and leftward motion with a speed of approximately 1-pixel per frame, and regions that are flickering or moving faster than the motion tuned channels.

feature vector) to be matched to correspondingly represented videos.

As an example, Fig. 2 illustrates the following five-way spatiotemporal oriented energy decomposition of a jumping jack action: unstructured (normalized ϵ), static (stationary texture), rightward and leftward motion (speed of 1 pixel/frame), and flicker/infinite motion (orientation orthogonal to the temporal axis); although, due to the relatively broad tuning of the filters employed, responses arise to a range of orientations about the peak tunings. In the initial frame, the stationary person has a large response in the static-related orientation channel and the background has a generally large response in the normalized ϵ channel due to its weak texture structure. During the action, the leg regions exhibit temporally asynchronous large responses in the orientation channels tuned to leftward and rightward motion due to the in and out movement of the legs. The arm regions exhibit a large response in the flicker channel due to the rapid up and down arm movements. Considered in unison, the spatiotemporal organization of these distributed oriented energy responses provide an action signature.

The constructed representation enjoys a number of attributes that are worth emphasizing.

1. Owing to the bandpass nature of the Gaussian derivative filters used in (1), the representation is invariant to additive photometric bias in the input signal. This follows from the fact that an additive bias manifests as a DC response that the front-end bandpass linear filters ignore.
2. Owing to the divisive normalization step, (5), the representation is invariant to multiplicative photometric bias. Specifically, the multiplicative bias appears in each of the unnormalized energy terms of the numerator and denominator of (5) and simply cancel out through division.
3. Owing to the marginalization step, (4), the representation is invariant to changes in appearance manifest as spatial orientation variation.

Overall, these three properties result in a robust pattern description that is invariant to changes unrelated to dynamic variation (e.g., different clothing), even while making explicit local orientation structure that arises with temporal variation (single motion, multiple motion, temporal flicker, etc.). In addition:

4. Owing to the oriented energies being defined over a spatiotemporal support region, (1), the representation can handle input data that are not exactly spatiotemporally aligned. This point is illustrated in Fig. 2, where it is seen that the energy responses are spatially broad about their local maxima; correspondingly, exact alignment between template and search video is not critical in matching: Matches can be driven by the broader responses, not just punctate peaks in the filter outputs.
5. Owing to the distributed nature of the representation, foreground clutter can be accommodated: Both the desirable action pattern structure and the undesirable clutter structure can be captured jointly so that the desirable components remain available for matching even in the presence of clutter. The approach's resilience to foreground clutter is demonstrated empirically in Section 3.1 (Fig. 3) through two examples, namely, dappled lighting and a thin fragmented occluder (i.e., a fence) superimposed over the actions.
6. The representation is efficiently realized via linear (separable convolution, pointwise addition) and pointwise nonlinear (squaring, division) operations [64], [65]; Section 2.4 provides a description of a real-time GPU implementation.

2.2 Spacetime Template Matching

To detect actions (as defined by a small template video) in a larger search video, the search video is scanned over all spacetime positions by sliding a 3D template over every spacetime position. At each position, the similarity between the oriented energy distributions (histograms) at the corresponding positions of the template and search volumes are computed.

To obtain a global match measure, $\Gamma(\mathbf{x})$, between the template and search videos at each image position, $\mathbf{x}$, of the search volume, the individual histogram similarity measurements are summed across the template

$$\Gamma(\mathbf{x}) = \sum_{\mathbf{u}} \gamma[\mathbf{S}(\mathbf{u}), \mathbf{T}(\mathbf{u}-\mathbf{x})], \tag{7}$$

where $\mathbf{u} = (u, v, w)$ ranges over the spacetime support of the template volume and $\gamma[\mathbf{S}(\mathbf{u}), \mathbf{T}(\mathbf{u}-\mathbf{x})]$ is the similarity between local distributions of the template, $\mathbf{T}$, and the search, $\mathbf{S}$, volumes. The peaks of the global similarity measure across the search volume represent potential match locations.

There are several histogram similarity measures that could be used [66]. Here, the Bhattacharyya coefficient [67] is used, as it takes into account the summed unity structure of distributions (unlike L_p-based match measures) and

yields an efficient implementation. The Bhattacharyya coefficient for two histograms $\mathbf{P}$ and $\mathbf{Q}$, each with B bins, is defined as

$$\gamma(\mathbf{P},\mathbf{Q}) = \sum_{b=1}^{B} \sqrt{P_b Q_b}, \tag{8}$$

with b the bin index. This measure is bounded below by zero and above by one [68], with zero indicating a complete mismatch, intermediate values indicating greater similarity, and one complete agreement. Significantly, the bounded nature of the Bhattacharyya coefficient makes it robust to modest amounts of outliers (e.g., as might arise during occlusion in the present application).

The final step consists of identifying peaks in the similarity volume, Γ, where peaks correspond to volumetric regions in the search volume that match closely with the template dynamics. For action spotting, the local maxima are identified in an iterative manner to avoid multiple detections around peaks: In the spacetime volumetric region about each peak the match score is suppressed (set to zero); this nonmaxima suppression process repeats until all remaining match scores are below a threshold, τ. In the experiments, the volumetric region of the template centered at the peak is used for suppression. For action recognition, only the global maximum is recovered.

Depending on the task, it may be desirable to weight the contribution of various regions in the template differently. For example, one may want to emphasize certain spatial regions and/or frames in the template. This can be accommodated with the following modification to the global match measure, (7):

$$\Gamma(\mathbf{x}) = \sum_{\mathbf{u}} \mathbf{w}(\mathbf{u})\gamma[\mathbf{S}(\mathbf{u}), \mathbf{T}(\mathbf{u}-\mathbf{x})], \tag{9}$$

where $\mathbf{w}$ denotes the weighting function. In some applications, it may also be desired to emphasize the contribution of certain dynamics in the template over others. For example, one may want to emphasize the dynamic over the unstructured and static information. This can be done by setting the weight in the match measure, (9), to $\mathbf{w} = 1 - (\hat{E}_\epsilon + \hat{E}_{\text{static}})$, with the oriented energy measure $\hat{E}_{\text{static}}$, (5), corresponding to static structure, i.e., nonmoving/zero-velocity, and $\hat{E}_\epsilon$ capturing local lack of structure, (6). An advantage of the developed representation is that it makes these types of semantically meaningful dynamics directly accessible.

For efficient search, one could resort to

1. spatiotemporal coarse-to-fine search [52], [7],
2. evaluation of the template on a coarser sampling of positions in the search volume,
3. evaluation of a subset of distributions in the template, and
4. early termination of match computation [69].

A drawback of these optimization strategies is that the target may be missed entirely. In this section, it is shown that exhaustive computation of the search measure, (7), can be realized in a computationally efficient manner.

Inserting the Bhattacharyya coefficient, (8), into the global match measure, (7), and reorganizing by swapping the spacetime and bin summation orders reveals that the expression is equivalent to the sum of cross-correlations between the individual bin volumes:

$$\Gamma(\mathbf{x}) = \sum_b \sum_{\mathbf{u}} \sqrt{S_b(\mathbf{u})}\sqrt{T_b(\mathbf{u}-\mathbf{x})} = \sum_b \sqrt{S_b} \star \sqrt{T_b}, \tag{10}$$

with $\star$ denoting correlation, b indexing histogram bins, and $\mathbf{u} = (u, v, w)$ ranging over template support.

Consequently, the correlation surface can be computed efficiently in the frequency domain using the convolution theorem of the Fourier transform [70], where the expensive correlation operations in spacetime are exchanged for relatively inexpensive pointwise multiplications in the frequency domain:

$$\Gamma(\mathbf{x}) = \mathcal{F}^{-1}\left\{\sum_b \mathcal{F}\{\sqrt{S_b}\}\mathcal{F}\left\{\sqrt{T_b'}\right\}\right\}, \tag{11}$$

with $\mathcal{F}\{\cdot\}$ and $\mathcal{F}^{-1}\{\cdot\}$ denoting the Fourier transform and its inverse, respectively, and T_b' the reflected template. In implementation, the Fourier transforms are realized efficiently by the fast Fourier transform (FFT).

2.3 Computational Complexity Analysis

Let $W_{\{\mathbf{T},\mathbf{S}\}}$, $H_{\{\mathbf{T},\mathbf{S}\}}$, and $D_{\{\mathbf{T},\mathbf{S}\}}$ be the width, height, and temporal duration, respectively, of the template, $\mathbf{T}$, and the search video, $\mathbf{S}$, and B denote the number of spacetime orientation histogram bins. The complexity of the correlation-based scheme in the spacetime domain, (10), is $O(B\prod_{i\in\{\mathbf{T},\mathbf{S}\}} W_i H_i D_i)$. In the case of the frequency domain-based correlation, (11), the 3D FFT can be realized efficiently by a set of 1D FFTs due to the separability of the kernel [61]. The computational complexity of the frequency domain-based correlation is $O[BW_{\mathbf{S}}H_{\mathbf{S}}D_{\mathbf{S}}(\log_2 D_{\mathbf{S}} + \log_2 W_{\mathbf{S}} + \log_2 H_{\mathbf{S}})]$.

In practice, on a standard CPU the overall runtime to compute the entire similarity volume between a $50 \times 25 \times 20$ template and a $144 \times 180 \times 200$ search video with six spacetime orientations and ϵ is 20 seconds (i.e., 10 frames/second) using the frequency-based scheme, (11), with the computation of the representation (Section 2.1) taking 16 seconds of the total time. In contrast, search in the spacetime domain, (10), takes 26 minutes. These timings are based on unoptimized Matlab code executing on a 2.3 GHz processor. In comparison, using the same sized input and a Pentium 3.0 GHz processor, [52] reports that this approach takes 30 minutes for exhaustive search.

Depending on the target application, additional savings of the proposed approach can be achieved by precomputing the search target representation offline. Also, since the representation construction and matching are highly parallelizable, real to near-real-time performance is possible through the use of widely available hardware and instruction sets, e.g., multicore CPUs, GPUs, and SIMD instruction sets. The following section describes a real-time GPU-based implementation of the action spotting approach.

2.4 Real-Time Implementation

In addition to its low computational complexity, the various processing stages of the proposed search approach are amenable to parallelization. In this section, details of a

TABLE 1
Running Times for the Three Major Parts of the Proposed GPU Implementation

GPU Model	Feature Extraction	Compute Similarity Volume	Find Peaks
GeForce 8400 GT {8}	60.3ms (32%)	123.0ms (66%)	3.1ms (2%)
Quadro FX 1800 {64}	13.7ms (35%)	24.1ms (62%)	1.0ms (3%)
GeForce 360M GTS {96}	7.6ms (19%)	30.3ms (77%)	1.6ms (4%)
GeForce 9800 GT {112}	9.5ms (33%)	18.1ms (64%)	0.8ms (3%)
GeForce 285 GTX {240}	3.7ms (25%)	10.6ms (71%)	0.6ms (4%)

Running times are given in milliseconds (ms) and the percentage of overall execution given in parentheses. Numbers in braces indicate the GPU cores available.

real-time GPU-based instantiation of the proposed action spotting approach for live video processing are provided. To realize the live action spotting approach, nVidia GPUs and the CUDA programming model [71] were chosen due to their relative maturity and ease of programming.

The efficient match algorithm described in Section 2.2 is founded on the assumption that the entire search volume is available. Consequently, this allows the computation of the 3D spacetime orientation features and the FFT-based matching steps to be performed in a single pass over the data. To accommodate causal processing for live videos, incoming frames must be processed, while results from previous frames are removed to avoid indefinite memory growth. This requirement can be achieved by using a "sliding temporal window," realized as a circular array of size T, where each incoming frame at time t replaces the oldest encountered frame at $t-T$. While this temporal windowing scheme precludes application of FFT-based convolution along the temporal dimension, efficient frame-wise computation can still be accomplished by realizing the xy space convolution via pointwise multiplication in the Fourier domain followed by pointwise addition along the t-dimension (i.e., a standard convolution operation along the temporal axis). The inverse FFT of the final result yields the desired slice of the similarity volume. The forward and inverse transforms are realized using CUDA's CUFFT library. The final peak finding step is highly parallelizable, where it can be realized by a binary tree-like reduction operation [72].

Table 1 gives the runtimes for each of the three major components of the algorithm executed on a live search video of size 320×184 and a template of size $64 \times 43 \times 30$ using a 7D histogram (six orientations and ϵ) for a number of nVidia cards that primarily differ in the number of GPU cores available. Indeed, oriented energies and similarity scores can be computed very efficiently. Note the performance of peak finding ultimately depends on the application and its choice can also be dictated by the architecture.

3 Empirical Evaluation

3.1 Action Spotting

The performance of the proposed action spotting algorithm has been evaluated on an illustrative set of test sequences. Matches are represented as a series of (spatial) bounding boxes that spatiotemporally outline the detected action. To realize the spatiotemporal orientation features for action spotting, six different spacetime orientations are made explicit, corresponding to static (no motion/orientation orthogonal to the image plane), leftward, rightward, upward, downward motion (1 pixel/frame movement), and flicker/infinite motion (orientation orthogonal to the temporal axis).

Fig. 3. Foreground clutter. Top three rows: Sample frames for walking left, one- and two-handed wave query templates. Fourth row: Sample walking left and one-handed wave detection results with foreground clutter in the form of local lighting variation caused by overhead dappled sunlight. Bottom row: Two-handed wave detections as action is performed beside and behind a chain link fence. Foreground clutter takes the form of superimposed static structure when the action is performed behind the fence.

3.1.1 Foreground Clutter

Fig. 3 shows results of the proposed approach on two outdoor scenes containing distinct forms of foreground clutter, which superimpose significant unmodeled patterning over the depicted actions and thereby test robustness to irrelevant structure in matching. The first example contains foreground clutter in the form of dappled sunlight with the query actions of walking left and one-handed wave. The second example contains foreground clutter in the form of superimposed static structure (i.e., pseudotransparency) caused by the chain-linked fence and the query action of two-handed wave. The first and second examples contain 365 and 699 frames, respectively, with the same spatial resolution of 321×185 pixels. All actions are spotted correctly; there are no false positives.

To quantify action spotting performance based on a single action exemplar, experiments were conducted on the CMU [20] and MSR [10] action datasets.

3.1.2 CMU Evaluation

The CMU action dataset [20] is comprised of five action categories, namely, pick-up, jumping jacks, push elevator button, one-handed wave, and two-handed wave, where each action is represented by a single training exemplar. The CMU dataset was captured with a handheld video camcorder in crowded environments with moving people and cars in the background. Also, there are large variations

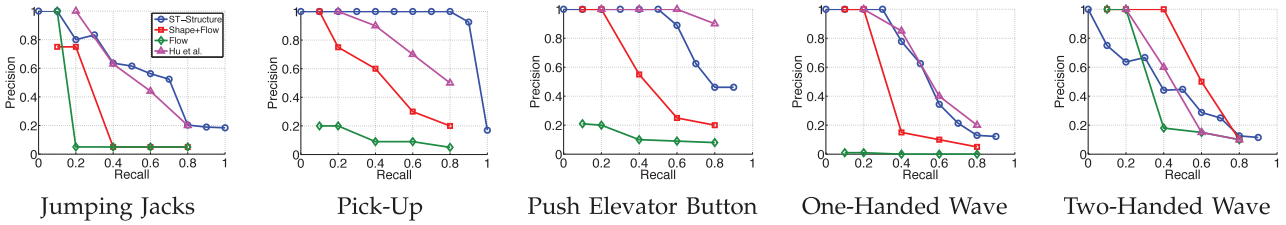


Fig. 4. Precision-Recall curves for the CMU action dataset. Precision-Recall plots: Blue curves correspond to the proposed approach, and red and green to baselines employing a single action exemplar, using shape plus flow with parts [20], or holistic flow [52], respectively (as reported in [20]), and purple to a recent discriminative training-based approach (motion plus appearance) [24].

in the performance of the actions, including their distance with respect to the camera.

Results are compared with ground truth labels included with the CMU dataset. The labels define the spacetime positions and extents of each action. For each action, a Precision-Recall (P-R) curve is generated by varying the similarity threshold from 0.6 to 0.97:

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{and} \quad \text{Recall} = \frac{TP}{nP}, \tag{12}$$

where TP is the number of true positives, FP is the number of false positives, and nP is the total number of positives in the dataset. In evaluation, the same testing protocol as in [20] is used. A detected action is considered a true positive if it has a (spacetime) volumetric overlap greater than 50 percent with a ground truth label. The same action templates from [20] are used, including the provided spacetime binary masks to emphasize the action over the background.

To gauge performance, the results of the proposed approach were compared to three recent action spotting approaches: 1) holistic flow [52], 2) parts-based shape plus flow [20], and 3) motion plus spatial appearance [24]. Similar to the proposed approach, the holistic flow approach only considers image dynamics, whereas the parts-based shape plus flow approach combines the same holistic flow approach in [52] with additional shape information recovered from a spatiotemporal over-segmentation procedure and embeds the matching process in a parts-based framework [73] to improve generalization. Furthermore, both approaches consider action spotting from a single action exemplar, as is the focus in the current approach. In contrast, the third approach [24] relies on a set of positive and negative action exemplars to train a discriminative classifier.

Fig. 4 shows P-R curves for each action. In comparison to the holistic flow approach [52], the proposed approach generally achieves superior spotting results, except in the case of two-handed wave; nevertheless, the proposed approach still outperforms holistic flow over most of the P-R plot in this case. Furthermore, as discussed in Section 2.3, the proposed approach is significantly superior from a computational complexity perspective. In comparison to the parts-based shape plus flow approach [20], the spotting performance of the proposed approach once again is generally superior, except in the case of two-handed wave. In both cases, two-handed wave is primarily confused with one-handed wave, resulting in a higher false positive rate and thus lower precision. Such confusions are not unreasonable given that the one-handed wave action is consistent with a significant portion of a two-handed wave. In contrast, the parts-based decomposition allows for greater flexibility in distinguishing one-handed versus two-handed wave. It should also be noted that the parts-based approach is computationally more expensive than the proposed approach since it combines the computationally expensive flow approach with a costly spacetime oversegmentation step. Even without the consideration of shape cues and a deformable model, the proposed approach generally outperforms [20], which demonstrates the discriminability of the proposed feature set focused solely on image dynamics and its robustness to a range of deformations.

In comparison to the discriminative learning motion plus appearance approach [24], performance is very similar for one-handed wave; for jumping jacks and two-handed wave, performance again is rather similar with some trend for the proposed approach to yield somewhat lower precision at low recall, albeit higher precision at higher recall. For push elevator button, performance is indistinguishable until the highest recall rates, where [24] performs better. For pick-up, the proposed approach performs significantly better across the entire operational range considered. In its instantiation of discriminative learning, the alternative approach makes use of one positive exemplar and instances of the remaining four actions as negative examples. In practice, reliance on discriminative learning may limit applicability to video browsing where a user might not provide negative or positive examples beyond the instance of concern.

3.1.3 MSR Evaluation

The recent MSR action dataset [10] contains the following three actions: boxing, hand clapping, and hand waving. The testing sequences are captured with cluttered and dynamic backgrounds, and the scale and performance of the actions vary significantly across the subjects. The training actions used with the MSR dataset are taken from the KTH action dataset [1]. This is an example of a cross-dataset experiment, where the training and test data are taken from different sources for the purpose of exercising the generalizability of an approach.

Action spotting performance of the proposed method was compared to the discriminative learning-based approach [10] introduced in conjunction with the MSR dataset. This learning-based approach described actions by a combination of features capturing spatial appearance (histogram of gradients) and motion (histogram of flow). Consistent with the current paper's goal of action spotting based on a *single* action template, three KTH templates per action were considered, one by one, for evaluating the proposed approach

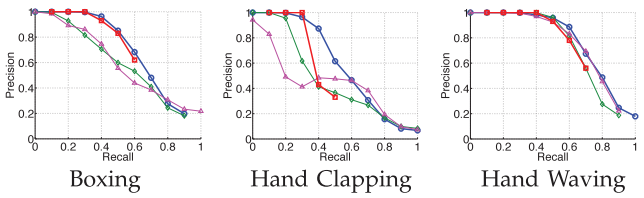


Fig. 5. Precision-Recall curves for the MSR action dataset. Blue, purple, and green curves correspond to the proposed approach while using three different templates and red to the baseline [10].

using the automatically generated crop windows provided elsewhere [8]. Furthermore, since boxing is a spatially asymmetric action, spotting was performed using the initial query template and its mirror-reflection. To generate the P-R curves for the boxing action, the detection windows realized independently with each template were jointly considered in the nonmaxima suppression step. As with the initial evaluation [10], a detected action is considered a true positive if the detection has a (spacetime) volumetric overlap greater than $1/8$ with the provided ground truth label.

Fig. 5 shows P-R curves for each action. In comparison to the baseline approach [10], the proposed approach generally achieves superior results across the entire P-R plot, especially in the high recall range, based on the best template for each action (blue curve). Results for the second and third best templates (purple and green curves, respectively) for the proposed approach are comparable with the baseline on the hand waving action and lower in the other actions. This plot of multiple P-R curves as a function of selected template also serves to indicate the sensitivity of the approach to template variation. Overall, even without consideration of appearance information and discriminative learning, where the majority of the nontarget actions present in the test set are used for training (as done in the baseline [10]), the proposed feature set proves to be efficacious in these cross-dataset comparisons.

3.2 Action Recognition

To quantify action recognition performance of the proposed approach, experiments were conducted on the standard KTH action [1] and UCF Sports [2] datasets. Further motivated by the current popularity of bag of words (BoW) approaches, an additional experiment was conducted with a correspondingly modified version of the proposed approach on the Hollywood2 dataset [3], as detailed below. For all experiments, the spatiotemporal orientation features were realized by making 21 different spacetime orientations explicit, corresponding to static (no motion/orientation orthogonal to the image plane), motion at two speeds (1 and 3 pixels/frame) and eight directions (leftward, rightward, upward, downward, and the intermediate diagonals), and four flicker/infinite motion-related orientations (horizontal, vertical, and diagonal orientations orthogonal to the temporal axis).

3.2.1 KTH Evaluation

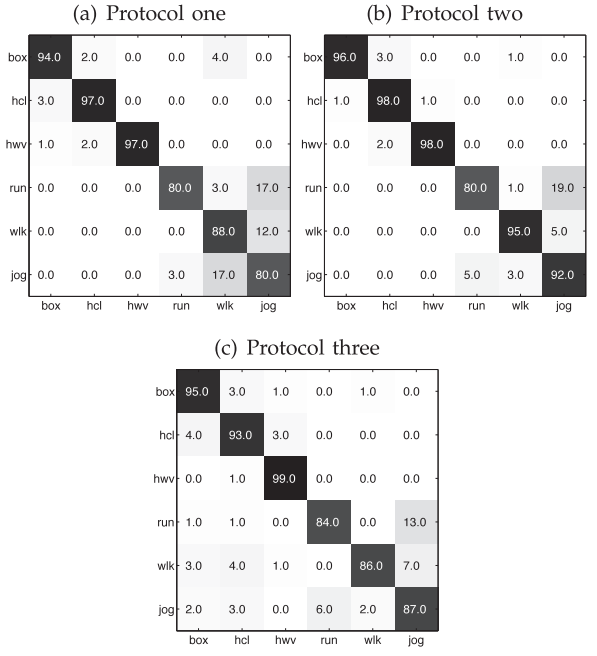
The KTH dataset contains the following six human actions: walking (wlk), jogging (jog), running (run), boxing (box), hand clapping (hcl), and hand waving (hwv). Each action is performed several times by 25 subjects under the following four different acquisition scenarios: scenario one (S1)—actions performed outdoors, scenario two (S2)—actions performed outdoors with scale variation, scenario three (S3)—actions performed outdoors with different clothes, and scenario four (S4)—actions performed indoors. Challenging aspects of this dataset include variations in human body shape, clothing, view angle, and spatiotemporal scale, which are not present in the Weizmann action dataset [19]. In the current evaluation, the same templates used in a previous paper [8] have been used. These templates are derived from an automatic spatiotemporal cropping process based on a saliency operator. Due to the automatic nature of the extraction process, the template extents are often inconsistent across instances of the same action and may introduce misclassifications due to the inexactness of the resulting crop windows.

To compare performance with others, experiments were conducted with the following three standard protocols. Protocol one uses nine subjects for testing and the remaining 16 subjects for training (i.e., eight subjects for training and eight subjects for validation); training and testing sequences used for the proposed approach are consistent with those specified in [1]. Protocol two is based on Leave-One-Out Cross Validation (LOOCV) [74]; for each run, sequences pertaining to one subject are chosen for testing and the sequences of the other 24 subjects as the training set; results are reported as the average of the 25 runs (i.e., each person takes a turn being "left-out"). Finally, protocol three uses 16 subjects that are randomly drawn for training and the remaining nine are used for testing. Recognition performance is reported as the average of five random splits. In protocols one and two, the scenarios are combined together and considered as a single large dataset. In protocol three, recognition is reported on a per scenario basis.

It is worth noting that within these testing protocols, different research groups have employed variations. As examples: Human tracking has been used to align the videos resulting in an actor-centric frame of [31], [47]; different subsets of actors in the training and testing sets from those of [1] have been used; subsets of the scenes have been used [29]; some groups report results based on the average classification rate of the individual scenarios treated independently, while others consider the various scenes as a single dataset (as done here for protocols one and two); the crop windows for the actions (where applicable) have not been consistent across evaluations; yet others pose the evaluation as a retrieval task and report results based on whether the correct detection is within a specified number of the Nearest Neighbors (NN) [7], [8]. Given these variations, comparative results presented in Table 3 have been grouped as accurately as possible according to the protocol that they most closely match.

In the proposed approach, classification of a given query is performed using nearest neighbor classification. Specifically, the label of the training set action yielding the highest peak similarity to the query is returned. During training, a greedy procedure is used to remove templates from the training set that yield high false positives as measured on the training set only. This training procedure is consistent with others that use a validation set to improve perfor-

TABLE 2
Confusion Matrix for the Three KTH Protocols



mance prior to classification. Based on this procedure, approximately 12 percent of the training templates were dropped from the training sets of each testing protocol.

The confusion matrices for the proposed approach for each testing protocol are shown in Tables 2a, 2b, and 2c. These matrices indicate that the proposed approach achieves generally high classification rates on a per action basis. Confusions predominately occur among the walking versus jogging and jogging versus running actions. These confusions can be accounted for by the high similarities of these class subsets and the relatively coarse sampling of the spatiotemporal orientation space used. A finer sampling of spatiotemporal orientation may improve the approach's ability to discriminate such highly similar classes.

Table 3 provides a comparison of the proposed approach with previous reported results in the literature under the various testing protocols; the table was reproduced from [7] and augmented with recent results in the literature. Under all three protocols, the proposed approach achieved superior recognition rates over the majority of previous methods, with those based on discriminative training (which is not performed in the current approach) yielding slightly better accuracy.

3.2.2 UCF Sports Evaluation

The UCF Sports dataset [2] contains the following 10 sports-related actions: diving (Dive), golf swinging (Golf), kicking (Kick), weightlifting (Lift), horse riding (Ride), running (Run), skateboarding (Skate), swinging on the pommel horse and on the floor (SwBen), swinging on the high bar (SwSide), and walking (Walk). The dataset contains 150 video sequences taken from actual sporting activities from a variety of sources, ranging from professional sports videography to amateur home video. Hand labeled spatial bounding boxes around the actions of interest are provided. To accommodate

TABLE 3
Comparison of Recognition Rates under Three Experimental Protocols with the KTH Action Dataset

Approaches	Protocol 1	Protocol 2	Protocol 3	Classifier
Proposed	89.34	93.18	90.51/98.12	NN/WT-3
[8]	82.7	–	–	NN
[51]	87.7	–	–	MIL
[7]	83.79	92.31	92.09	WT-3
[47]	–	(95.33)	–	NN
[33]	–	88.3	–	NN
[27]	–	94	–	SVM
[49]	(90.5)	–	–	AdaBoost
[35]	91.8	–	–	SVM
[36]	–	94.2	–	SVM
[34]	–	83.33	–	pLSA
[55]	–	–	91.7†	SVM
[21]	–	80.9	–	SVM
[31]	–	71.83 (91.6)	–	pLSA
[29]	–	81.17*	–	SVM
[40]	62.96	–	–	cascade
[1]	71.72	–	–	SVM

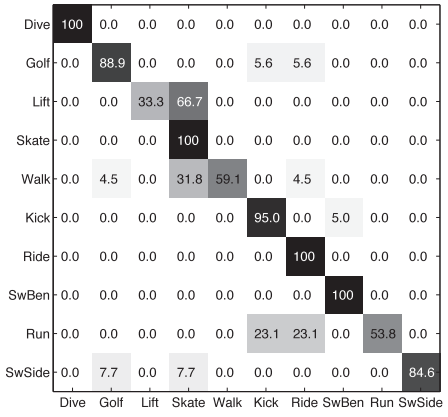
*Parentheses indicate that results were obtained with video sequences segmented and aligned by hand. * indicates only a subset of the dataset was used (i.e., S1 and S3). † indicates that background subtraction was used to delineate the target region prior to classification. WT-3 indicates that a correct recognition was determined if the correct detection was within the top three nearest detections.*

the sliding volume nature of the proposed approach, temporal windows were introduced that capture approximately a single cycle of the action of interest; these temporal windows were typically around 30 frames in length.

Consistent with previous reported evaluations on the UCF dataset [2], [43], [44], [45], [46], evaluation of the current approach was performed using LOOCV. The majority of these previous approaches [43], [44], [45], [46] rely on a global bag-of-words (GBoW) representation and thus may benefit from contextual cues from the background. To measure the impact of background-related contextual cues, two additional baseline methods were evaluated. The first baseline consisted of a global bag-of-words video representation (cf. [44]). The second baseline was similar to the first with the exception that the words within each spatial bounding box outlining the action (provided with the dataset) were removed from consideration and thus performance is solely driven by the background bag-of-words (BBoW), i.e., the context. In both cases, a histogram of 3D gradient features that jointly capture appearance and dynamic information was extracted densely across each video using code and parameter settings (i.e., 4,000 visual words) from a recent evaluation of action recognition approaches [44]. Classification for these two baselines was realized using a linear SVM classifier [75] with the regularization weighting $C = 1{,}000$. Note that the approach proposed in this paper is exposed to only a very limited amount of background context due to the tight nature of the provided (spatial) template bounding boxes. The proposed approach used the same NN classification procedure described previously in Section 3.2.1.

The confusion matrix for the proposed approach is shown in Table 4. Performance on a per-class basis is generally high, with difficulties predominately isolated to the lift, run and walk actions. Table 5 provides a comparison of the proposed approach with previous reported results in the literature and the two global bag-of-words baselines.

TABLE 4
Confusion Matrix for the UCF Sports Data Set



Compared to the lone approach from the baseline set formulated as a sliding volume [2], and therefore most comparable to the proposed approach, significant improvement in accuracy is observed. While some of the state-of-the-art global bag-of-word approaches achieve somewhat higher performance [44], [45], [46], it is interesting to cast their performance in the light of the GBoW/BBoW comparison: Here, it is seen that consideration of the background only (BBoW) can achieve comparable performance to action plus background (GBoW) for global bag-of-words approaches, which raises the question of whether such approaches are recognizing the background or the action per se. In contrast, the proposed approach achieved a high accuracy, even given its simple NN classifier and very limited exposure to background context.

3.2.3 Hollywood2 Evaluation

The Hollywood2 dataset [3] contains the following 12 actions: answering the phone, driving the car, eating, fighting, getting out of the car, hand shaking, hugging, kissing, running, sitting down, sitting up, and standing up. In total, there are 1,707 videos divided into a training and testing set consisting of 823 and 884 sequences, respectively.

In this experiment, each video was represented as a bag of words. To make direct comparison with a wide range of popular BoW-related work, the protocol from a recent action recognition evaluation of feature descriptors was followed [44]. Specifically, for each video, the spacetime oriented features, (5) and (6), were first densely sampled uniformly across spacetime with a sampling rate of 5 pixels and frames in the spatial and temporal dimensions, respectively. Next, each descriptor was formed by concatenating spacetime neighbouring oriented features in a $3 \times 3 \times 3$ grid and mapping the result to its nearest visual word. Finally, each video was represented by the histogram of frequencies for each word. The visual word vocabulary was constructed with K-means clustering, where $K = 4{,}000$. To speed up the vocabulary construction, a subset of 100,000 randomly selected training features was used for clustering. In training and testing, features were assigned to their closest vocabulary word based on the euclidean distance. For classification, a nonlinear support vector machine (SVM) [76] with a chi-square kernel was used [35].

TABLE 5
Comparison of Average Per-Class Recognition Rates with LOOCV Using the UCF Data Set

Proposed	[2]	[43]	[44]	[45]	[46]	GBoW	BBoW
81.5	69.2	79.3	85.6	86.7	87.3	81.9	80.7

TABLE 6
Summary of Average Precision Rates on Hollywood2

AnswerPhone	DriveCar	Eat	FightPerson	GetOutCar	HandShake	HugPerson	Kiss	Run	SitDown	SitUp	StandUp	Average
.22	.83	.54	.72	.32	.16	.37	.59	.76	.56	.18	.56	.48

Table 6 provides a class-wise summary of the average precision rates on Hollywood2. Table 7 provides a comparison of the proposed BoW model with several popular feature descriptors computed densely (for implementation details, see [44]): spatial histogram of gradients (HoG) [35], histogram of optical flow (HoF) [35], HoG and HoF concatenated (HoG/HoF), and spacetime histogram of gradients (HoG3D) [56]. Here, results are reported only for dense feature recovery because a recent empirical comparison [44] showed that performance with dense sampling equaled or exceeded sparse recovery for all considered features. Among the various single descriptors, HoG, HoF, and HoG3D, and the concatenated feature, HoG/HoF, the proposed descriptor performed best. Interestingly, unlike the features compared, the proposed descriptor is currently constructed at a single scale only. A similar multiscale extension to the proposed descriptor may provide additional performance gains.

4 Discussion and Summary

The main contribution of this paper is the representation of visual spacetime via spatiotemporal orientation distributions for the purpose of computationally efficient action spotting and recognition. It has been shown that this tack can accommodate variable appearance of the same action, rapid dynamics, multiple actions in the field-of-view, and is robust to scene clutter (both foreground and background) while being amenable to efficient computation.

A current limitation of the proposed action spotting approach is related to the use of a single monolithic template for any given action. This choice limits the ability to generalize across the range of observed variability as an

TABLE 7
Mean Average Precision Rate (Percent) for Various Descriptors Computed Densely on the Hollywood2 Data Set

Proposed	HoG [35]	HoF [35]	HoG/HoF [35]	HoG3D [56]
48.4	39.4	45.5	47.4	45.3

Results related to HoG, HoF, and HoG3D features are reproduced from [44]. All results are based on a dictionary of 4,000 words.

action is depicted in natural conditions, such as large spatial and temporal deformations due to anthropomorphic attributes (e.g., height and shape) and action execution speed, respectively. Further response to these action variations motivates future investigation of deformable action templates (e.g., parts-based), which allows for flexibility in matching template (sub)components, cf. [20]. More general reduction in false positives should be achievable through incorporation of complementary cues into the action template description, e.g., shape. Along these lines, it is interesting to note that the presented empirical results attest that the proposed approach already is robust to a range of deformations between the template and search target (e.g., modest changes in spatial scale, rotation, and execution speed as well as individual performance nuances). Such robustness is due to the relatively broad tuning of the oriented energy filters, which discount minor differences in spacetime orientations between template and target.

In summary, this paper has presented novel approaches to action spotting and recognition. The approaches are founded on a distributed characterization of visual spacetime in terms of 3D, (x, y, t), spatiotemporal orientation that captures underlying pattern dynamics. For both approaches, no elaborate training is introduced; in the case of action spotting, results are shown on single query videos, whereas, for action recognition, simple nearest neighbors or a standard SVM formulation is used. In addition, details of a real-time implementation have been provided. Empirical evaluation on a challenging set of image sequences, including quantitative comparisons on standard datasets, demonstrates the potential of the proposed approach.

ACKNOWLEDGMENTS

Portions of this research were funded by an NSERC Discovery Grant to R. Wildes. The authors thank H. Seo and P. Milanfar for providing their cropped KTH templates, and T. Lian for helpful discussion.

REFERENCES

[1] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing Human Actions: A Local SVM Approach," *Proc. 17th Int'l Conf. Pattern Recognition,* pp. 32-36, 2004.

[2] M. Rodriguez, J. Ahmed, and M. Shah, "Action MACH a Spatio-Temporal Maximum Average Correlation Height Filter for Action Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2008.

[3] M. Marszalek, I. Laptev, and C. Schmid, "Actions in Context," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 2929-2936, 2009.

[4] K. Derpanis and R. Wildes, "The Structure of Multiplicative Motions in Natural Imagery," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 32, no. 7, pp. 1310-1316, July 2010.

[5] J. Hays and A. Efros, "Scene Completion Using Millions of Photographs," *Proc. ACM Siggraph,* vol. 26, no. 1, 2007.

[6] A. Torralba, R. Fergus, and W. Freeman, "80 Million Tiny Images: A Large Database for Non-Parametric Object and Scene Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 30, no. 11, pp. 1958-1970, Nov. 2008.

[7] H. Ning, T. Han, D. Walther, M. Liu, and T. Huang, "Hierarchical Space-Time Model Enabling Efficient Search for Human Actions," *IEEE Trans. Circuits and Systems for Video Technology,* vol. 19, no. 6, pp. 808-820, June 2009.

[8] H. Seo and P. Milanfar, "Action Recognition from One Example," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 33, no. 5, pp. 867-882, May 2011.

[9] K. Derpanis, M. Sizintsev, K. Cannons, and R. Wildes, "Efficient Action Spotting Based on a Spacetime Oriented Structure Representation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2010.

[10] J. Yuan, Z. Liu, and Y. Wu, "Discriminative Video Pattern Search for Efficient Action Detection," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 33, no. 9, pp. 1728-1743, Sept. 2011.

[11] P. Turaga, R. Chellappa, V. Subrahmanian, and O. Udrea, "Machine Recognition of Human Activities: A Survey," *IEEE Trans. Circuits and Systems for Video Technology,* vol. 18, no. 11, pp. 1473-1488, Nov. 2008.

[12] R. Poppe, "A Survey on Vision-Based Human Action Recognition," *Image and Vision Computing,* vol. 28, no. 6, pp. 976-990, 2010.

[13] Y. Yacoob and M. Black, "Parameterized Modeling and Recognition of Activities," *Computer Vision and Image Understanding,* vol. 73, no. 2, pp. 232-247, 1999.

[14] D. Ramanan and D. Forsyth, "Automatic Annotation of Everyday Movements," *Proc. Neural Information Processing Systems,* 2003.

[15] C. Fanti, L. Zelnik Manor, and P. Perona, "Hybrid Models for Human Motion Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 1166-1173, 2005.

[16] S. Ali, A. Basharat, and M. Shah, "Chaotic Invariants for Human Action Recognition," *Proc. 11th IEEE Int'l Conf. Computer Vision,* 2007.

[17] A. Bobick and J. Davis, "The Recognition of Human Movement Using Temporal Templates," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 23, no. 3, pp. 257-267, Mar. 2001.

[18] D. Weinland, R. Ronfard, and E. Boyer, "Free Viewpoint Action Recognition Using Motion History Volumes," *Computer Vision and Image Understanding,* vol. 103, nos. 2/3, pp. 249-257, 2006.

[19] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as Space-Time Shapes," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 29, no. 12, pp. 2247-2253, Dec. 2007.

[20] Y. Ke, R. Sukthankar, and M. Hebert, "Event Detection in Crowded Videos," *Proc. 11th IEEE Int'l Conf. Computer Vision,* 2007.

[21] Y. Ke, R. Sukthankar, and M. Hebert, "Spatio-Temporal Shape and Flow Correlation for Action Recognition," *Proc. WVS,* 2007.

[22] A. Yilmaz and M. Shah, "A Differential Geometric Approach to Representing the Human Actions," *Computer Vision and Image Understanding,* vol. 109, no. 3, pp. 335-351, 2008.

[23] Z. Zhang, Y. Hu, S. Chan, and L. Chia, "Motion Context: A New Representation for Human Action Recognition," *Proc. 10th European Conf. Computer Vision,* pp. 817-829, 2008.

[24] Y. Hu, L. Cao, F. Lv, S. Yan, Y. Gong, and T. Huang, "Action Detection in Complex Scenes with Spatial and Temporal Ambiguities," *Proc. 12th IEEE Int'l Conf. Computer Vision,* pp. 128-135, 2009.

[25] T. Kobayashi and N. Otsu, "Three-Way Auto-Correlation Approach to Motion Recognition," *Pattern Recognition Letters,* vol. 30, no. 3, pp. 212-221, 2009.

[26] Z. Lin, Z. Jiang, and L. Davis, "Recognizing Actions by Shape-Motion Prototype Trees," *Proc. 12th IEEE Int'l Conf. Computer Vision,* pp. 444-451, 2009.

[27] X. Sun, M. Chen, and A. Hauptmann, "Action Recognition via Local Descriptors and Holistic Features," *Proc. IEEE CVPR Workshop for Human Communicative Behavior Analysis,* pp. 58-65, 2009.

[28] I. Laptev, "On Space-Time Interest Points," *Int'l J. Computer Vision,* vol. 64, nos. 2/3, pp. 107-123, 2005.

[29] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior Recognition via Sparse Spatio-Temporal Features," *Proc. Second Joint IEEE Int'l Workshop Visual Surveillance and Performance Evaluation of Tracking and Surveillance,* pp. 65-72, 2005.

[30] A. Oikonomopoulos, I. Patras, and M. Pantic, "Spatiotemporal Salient Points for Visual Recognition of Human Actions," *IEEE Trans. Systems, Man, and Cybernetics Part B: Cybernetics,* vol. 36, no. 3, pp. 710-719, June 2005.

[31] S. Wong, T. Kim, and R. Cipolla, "Learning Motion Categories Using Both Semantic and Structural Information," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2007.

[32] G. Willems, T. Tuytelaars, and L. Van Gool, "An Efficient Dense and Scale-Invariant Spatio-Temporal Interest Point Detector," *Proc. 10th European Conf. Computer Vision,* pp. 650-666, 2008.

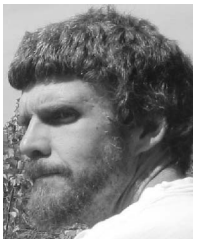
[33] K. Rapantzikos, Y. Avrithis, and S. Kollias, "Dense Saliency-Based Spatiotemporal Feature Points for Action Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2009.

[34] J. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words," *Int'l J. Computer Vision,* vol. 79, no. 3, pp. 299-318, 2008.
[35] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning Realistic Human Actions from Movies," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2008.
[36] J. Liu and M. Shah, "Learning Human Actions via Information Maximization," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2008.
[37] K. Mikolajczyk and H. Uemura, "Action Recognition with Motion-Appearance Vocabulary Forest," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2008.
[38] J. Liu, J. Luo, and M. Shah, "Recognizing Realistic Actions from Videos 'in the Wild'," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2009.
[39] J. Sun, X. Wu, S. Yan, L. Cheong, T. Chua, and J. Li, "Hierarchical Spatio-Temporal Context Modeling for Action Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2009.
[40] Y. Ke, R. Sukthankar, and M. Hebert, "Efficient Visual Event Detection Using Volumetric Features," *Proc. 10th IEEE Int'l Conf. Computer Vision,* pp. 166-173, 2005.
[41] N. Apostoloff and A. Fitzgibbon, "Learning Spatiotemporal T-Junctions for Occlusion Detection," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 553-559, 2005.
[42] A. Gilbert, J. Illingworth, and R. Bowden, "Action Recognition Using Mined Hierarchical Compound Features," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 33, no. 1, pp. 883-897, May 2011.
[43] L. Yeffet and L. Wolf, "Local Trinary Patterns for Human Action Recognition," *Proc. 12th IEEE Int'l Conf. Computer Vision,* pp. 492-497, 2009.
[44] H. Wang, M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of Local Spatio-Temporal Features for Action Recognition," *Proc. British Machine Vision Conf.,* 2009.
[45] A. Klaser, "Learning Human Actions in Videos," PhD dissertation, Université de Grenoble, 2010.
[46] A. Kovashka and K. Grauman, "Learning a Hierarchy of Discriminative Space-Time Neighborhood Features for Human Action Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 2046-2053, 2010.
[47] T. Kim and R. Cipolla, "Canonical Correlation Analysis of Video Volume Tensors for Action Categorization and Detection," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 31, no. 8, pp. 1415-1428, Aug. 2009.
[48] A. Efros, A. Berg, G. Mori, and J. Malik, "Recognizing Action at a Distance," *Proc. Ninth IEEE Int'l Conf. Computer Vision,* pp. 726-733, 2003.
[49] A. Fathi and G. Mori, "Action Recognition by Learning Mid-Level Motion Features," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2008.
[50] C. Yeo, P. Ahammad, K. Ramchandran, and S. Sastry, "High-Speed Action Recognition and Localization in Compressed Domain Videos," *IEEE Trans. Circuits and Systems for Video Technology,* vol. 18, no. 8, pp. 1006-1015, Aug. 2008.
[51] S. Ali and M. Shah, "Human Action Recognition in Videos Using Kinematic Features and Multiple Instance Learning," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 32, no. 2, pp. 288-303, Feb. 2010.
[52] E. Shechtman and M. Irani, "Space-Time Behavior-Based Correlation—OR—How to Tell If Two Underlying Motion Fields Are Similar without Computing Them?" *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 29, no. 11, pp. 2045-2056, Nov. 2007.
[53] P. Matikainen, R. Sukthankar, M. Hebert, and Y. Ke, "Fast Motion Consistency through Matrix Quantization," *Proc. British Machine Vision Conf.,* 2008.
[54] O. Chomat and J. Crowley, "Probabilistic Recognition of Activity Using Local Appearance," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 104-109, 1999.
[55] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, "A Biologically Inspired System for Action Recognition," *Proc. 11th IEEE Int'l Conf. Computer Vision,* 2007.
[56] A. Klaser, M. Marszalek, and C. Schmid, "A Spatio-Temporal Descriptor Based on 3D-Gradients," *Proc. British Machine Vision Conf.,* 2008.
[57] Google, "Image Search," http://images.google.com, 2012.
[58] D. Heeger, "Optical Flow from Spatiotemporal Filters," *Int'l J. Computer Vision,* vol. 1, no. 4, pp. 279-302, 1988.
[59] K. Derpanis and R. Wildes, "Early Spatiotemporal Grouping with a Distributed Oriented Energy Representation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2009.
[60] K. Derpanis and R. Wildes, "Dynamic Texture Recognition Based on Distributions of Spacetime Oriented Structure," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2010.
[61] B. Jähne, *Digital Image Processing,* sixth ed. Springer, 2005.
[62] A. Watson and A. Ahumada, "A Look at Motion in the Frequency Domain," *Proc. Motion Workshop,* pp. 1-10, 1983.
[63] E. Adelson and J. Bergen, "Spatiotemporal Energy Models for the Perception of Motion," *J. Optical Soc. Am.-A,* vol. 2, no. 2, pp. 284-299, 1985.
[64] W. Freeman and E. Adelson, "The Design and Use of Steerable Filters," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 13, no. 9, pp. 891-906, Sept. 1991.
[65] K. Derpanis and J. Gryn, "Three-Dimensional nth Derivative of Gaussian Separable Steerable Filters," *Proc. IEEE Int'l Conf. Image Processing,* pp. 553-556, 2005.
[66] Y. Rubner, J. Puzicha, C. Tomasi, and J. Buhmann, "Empirical Evaluation of Dissimilarity Measures for Color and Texture," *Computer Vision and Image Understanding,* vol. 84, no. 1, pp. 25-43, 2001.
[67] A. Bhattacharyya, "On a Measure of Divergence between Two Statistical Populations Defined by Their Probability Distribution," *Bull. Calcutta Math. Soc.,* vol. 35, pp. 99-110, 1943.
[68] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-Based Object Tracking," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 25, no. 5, pp. 564-577, May 2003.
[69] D. Barnea and H. Silverman, "A Class of Algorithms for Fast Digital Image Registration," *IEEE Trans. Computers,* vol. 21, no. 2, pp. 179-186, Feb. 1972.
[70] R. Bracewell, *The Fourier Transform and Its Applications.* McGraw-Hill, 2000.
[71] nVidia CUDA, www.nvidia.com/object/cuda_home.html, 2012.
[72] M. Harris, *Optimizing Parallel Reduction in CUDA,* NVIDIA Developer Technology, 2007.
[73] P. Felzenszwalb and D. Huttenlocher, "Pictorial Structures for Object Recognition," *Int'l J. Computer Vision,* vol. 61, no. 1, pp. 55-79, 2005.
[74] S. Theodoridis and K. Koutroumbas, *Pattern Recognition,* third ed. Academic Press, 2006.
[75] O. Chapelle, P. Haffner, and V. Vapnik, "Support Vector Machines for Histogram-Based Image Classification," *IEEE Trans. Neural Networks,* vol. 10, no. 5, pp. 1055-1064, Sept. 1999.
[76] C. Chang and C. Lin, "LIBSVM: A Library for Support Vector Machines," 2001.

Konstantinos G. Derpanis received the BSc (Honours) degree in computer science from the University of Toronto, Canada, in 2000, and the MSc and PhD degrees in computer science from York University, Canada, in 2003 and 2010, respectively. Subsequently, he was a postdoctoral researcher in the GRASP Laboratory at the University of Pennsylvania. In 2012, he joined the Department of Computer Science at Ryerson University, Toronto, where he is an assistant professor. His major research interests include motion analysis and human motion/gesture understanding. He is a member of the IEEE.

Mikhail Sizintsev received the BSc (Honours), MSc, and PhD degrees in computer science from York University, Canada, in 2004, 2006, and 2012, respectively. Currently, he is a research scientist at SRI International Sarnoff in Princeton, New Jersey. His major areas of research include stereo and motion with emphasis in spatiotemporal processing and analysis, visual navigation, and augmented reality. He is a member of the IEEE.

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



Richard P. Wildes received the PhD degree from the Massachusetts Institute of Technology in 1989. Subsequently, he joined Sarnoff Corporation in Princeton, New Jersey, as a member of the technical staff in the Vision Technologies Group. In 2001, he joined the Department of Computer Science and Engineering at York University, Toronto, where he is an associate professor and a member of the Centre for Vision Research. Honors include receiving a Sarnoff Corporation Technical Achievement Award, the IEEE D.G. Fink Prize Paper Award for his *Proceedings of the IEEE* publication "Iris Recognition: An Emerging Biometric Technology," and twice giving invited presentations to the US National Academy of Sciences. His main areas of research interest are computational vision, as well as allied aspects of image processing, robotics, and artificial intelligence. He is a member of the IEEE.