Efficient Action Spotting Based on a Spacetime Oriented Structure Representation

Konstantinos G. Derpanis, Mikhail Sizintsev, Kevin Cannons and Richard P. Wildes
Department of Computer Science and Engineering
York University
Toronto, Ontario, Canada
{kosta,sizints,kcannons,wildes}@cse.yorku.ca

Abstract

This paper addresses action spotting, the spatiotemporal detection and localization of human actions in video. A novel compact local descriptor of video dynamics in the context of action spotting 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. An important aspect of the descriptor is that it 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 across candidate video sequences. Empirical evaluation of the approach on a set of challenging natural videos suggests its efficacy.

1. Introduction

This paper addresses the problem of detecting and localizing spacetime patterns, as represented by a single query video, in a reference video database. Specifically, patterns of current concern are those induced by human actions. This problem is referred to as *action spotting*. The term "action" refers to a simple dynamic pattern executed by a person over a short duration of time (e.g., walking and hand waving). Potential applications of the presented approach include video indexing and browsing, surveillance, visually-guided interfaces and tracking initialization.

A key challenge in action spotting arises from the fact that the same underlying pattern dynamics can yield very different image intensities due to spatial appearance differences, as with changes in clothing and live action versus animated cartoon content. 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: Background clutter arises when actions are depicted in front of complicated, possibly dynamic, backdrops; foreground clutter arises when actions are depicted with distractions superimposed, as with dynamic lighting, pseudo-transparency (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

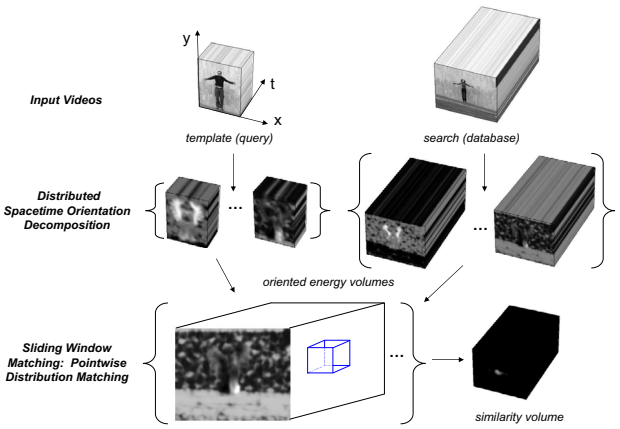

Figure 1. Overview of approach to action spotting. (top) A template (query) and search (database) video serve as input; both the template and search videos depict the action of "jumping jacks" taken from the Weizmann action data set [11]. (middle) Application of spacetime oriented energy filters decomposes the input videos into a distributed representation according to 3D, (x, y, t), spatiotemporal orientation. (bottom) In a sliding window manner, the distribution of oriented energies of the template is compared to the search distribution at corresponding positions to yield a similarity volume. Finally, significant local maxima in the similarity volume are identified.

possible to tease action information from clutter.

For 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 image dynamics (e.g., single motion, multiple superimposed motions, 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. Figure 1 provides an overview of the approach.

A wealth of work has considered the analysis of human actions from visual data [26]. A brief survey of representative approaches follows.

Tracking-based methods begin by tracking body parts and/or joints and classify actions based on features extracted from the motion trajectories (e.g., [28, 19, 1]). 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 [3, 11, 15]. 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 silhouette) and thus yield ambiguous information.

Recently, spacetime interest points have emerged as a popular means for action classification [22, 8, 17, 16]. Interest points typically are taken as spacetime loci that exhibit variation along all spatiotemporal dimensions and provide a sparse set of descriptors to guide action recognition. Sparsity is appealing as it yields significant reduction in computational effort; however, interest point detectors often fire erratically on shadows and highlights [14], and along object occluding boundaries, which casts doubt on their applicability to cluttered natural imagery. Additionally, for actions substantially comprised of smooth motion, important information is ignored in favour of a small number of possibly insufficient interest points.

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., optical flow, spatiotemporal gradients and other filter responses selective to both spatial and temporal orientation), typically matched to a video of interest in a sliding window formulation. Chief advantages of this framework include avoidance of problematic localization, tracking and segmentation preprocessing of the input video; however, such approaches can be computationally intensive. Further limitations are tied to the particulars of the image measurement used to define the template.

Optical flow-based methods (e.g., [9, 14, 20]) 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 [23, 15]. This tack yields a compact way to characterize visual spacetime locally, with template video matches 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., [5, 13, 18]) suffer from their inability to generalize across differences in spatial appearance (e.g., different clothing) of the same action.

The spacetime features used in the present work derive from filter techniques that capture dynamic aspects of visual spacetime with robustness to purely spatial appearance, as developed previously (e.g., in application to motion estimation [24] and video segmentation [7]). The present work appears to be the first to apply such filtering to action analysis. Other notable applications of similar spacetime oriented energy filters include, pattern categorization [27], tracking [4] and spacetime stereo [25].

In the light of previous work, the major contributions of the present paper are as follows. (i) A novel compact local oriented energy feature set is developed for action spotting. This representation supports fine delineations of visual spacetime structure to capture the rich underlying dynamics of an action from a single query video. (ii) 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 person localization, tracking, motion estimation, figure-ground segmentation or learning. (iii) 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. (iv) The approach is demonstrated on a set of challenging natural videos.

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, 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 with rapid actions) correspond to orientations that become orthogonal to the temporal axis. Further, appropriate definition of local spacetime oriented energy measurements can yield invariance to purely spatial pattern characteristics and support action spotting as an actor changes spatial appearance. Based on these observations, the developed action spotting approach makes use of such measurements as local features that are combined into spacetime templates to maintain relative geometric spacetime positions.

2.1. Features: Spacetime orientation

The desired spacetime 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 $\mathbf{x} = (x, y, t)$ spacetime position. The responses of the image data to this filter are pointwise rectified (squared) and integrated (summed) over a spacetime neighbourhood, Ω, to yield the following locally aggregated pointwise energy measurement

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

where $I \equiv I(\mathbf{x})$ denotes the input imagery and $*$ convolution. Notice that while the employed Gaussian derivative filter is phase-sensitive, summation over the support region ameliorates this sensitivity to yield a measurement of signal energy at orientation θ. More specifically, this follows from Rayleigh's theorem [12] that specifies the phase-independent signal energy in the frequency passband of the Gaussian derivative:

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

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 behaviour and thereby are 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 one wants to be invariant to appearance, while being sensitive to dynamic properties. This 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 spacetime orientation (e.g., image velocity) manifests itself as a plane through the origin in the frequency domain [12]. Correspondingly, summation across a set of x-y-t-oriented energy measurements consistent with a single frequency domain plane through the origin is indicative of energy along the associated spacetime 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 [10]. Let each plane be parameterized by its unit normal, $\hat{\mathbf{n}}$; a set of equally spaced $N + 1$ directions within the plane are given as

$$\hat{\theta}_i = \cos\left(\frac{2\pi i}{N+1}\right)\hat{\theta}_a(\hat{\mathbf{n}}) + \sin\left(\frac{2\pi i}{N+1}\right)\hat{\theta}_b(\hat{\mathbf{n}}), \qquad (3)$$

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 $0 \leq i \leq N$.

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

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

with $\hat{\theta}_i$ one of $N + 1 = 4$ specified directions, (3), and each $E_{\hat{\theta}_i}$ calculated via the oriented energy filtering, (1). In the present implementation, six different spacetime orientations are made explicit, corresponding to

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

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

static (no motion/orientation orthogonal to the image plane), leftward, rightward, upward, downward motion (one pixel/frame movement), 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.

Finally, the marginalized energy measurements, (4), are confounded by the local contrast of the signal and as a result increase monotonically with contrast. This makes it impossible to determine whether a high response for a particular spacetime 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 spacetime orientation, the energy measures are normalized by the sum of consort planar energy responses at each point,

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

where M denotes the number of spacetime 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 six oriented, appearance marginalized energy measurements, (4), Eq. (5) produces a corresponding set of six normalized, marginalized oriented energy measurements. To this set a seventh measurement is added that explicitly captures lack of structure via normalized ϵ,

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

to yield a seven dimensional feature vector at each point in the image data. (Note that 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.)

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 spacetime 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 7D orientation feature vector) to be matched to correspondingly represented videos where actions are to be spotted.

The constructed representation enjoys a number of attributes that are worth emphasizing. (i) Owing to the bandpass nature of the Gaussian derivative filters (1), the representation is invariant to additive photometric bias in the input signal. (ii) Owing to the divisive normalization (5), the representation is invariant to multiplicative photometric bias. (iii) Owing to the marginalization (4), the representation is invariant to changes in appearance manifest as spatial orientation variation. Overall, these three invariances result in a robust pattern description that is invariant to changes that do not correspond 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.). (iv) Owing to the oriented energies being defined over a spatiotemporal support region, (1), the representation can deal with input data that are not exactly spatiotemporally aligned. (v) 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. (vi) The representation is efficiently realized via linear (separable convolution, pointwise addition) and pointwise non-linear (squaring, division) operations [6].

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, $M(\mathbf{x})$, 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:

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

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

There are several histogram similarity measures that could be used [21]. Here, the Bhattacharyya coefficient [2] is used, as it takes into account the summed unity structure of distributions (unlike L_p-based match measures) and yields to efficient implementation (see Section 2.2.2). The Bhattacharyya coefficient for two histograms $\mathbf{P}$ and $\mathbf{Q}$, each with B bins, is defined as

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

with b the bin index. This measure is bounded below by zero and above by one, 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 small outliers (e.g., as might arise during occlusion in the present application).

The final step consists of identifying peaks in the similarity volume, M, where peaks correspond to volumetric regions in the search volume that match closely with the template dynamics. The local maxima in the volume are identified via non-maxima suppression. In the experiments, the volumetric region of the template centered at the peak is used for suppression.

2.2.1 Weighting template contributions

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:

$$M(\mathbf{x}) = \sum_{\mathbf{u}} \mathbf{w}(\mathbf{u}) m[\mathbf{S}(\mathbf{u}), \mathbf{T}(\mathbf{u}-\mathbf{x})], \qquad (9)$$

where $\mathbf{w}$ denotes the weighting function. In some scenarios, 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 $\hat{E}_{\text{static}}$ the oriented energy measure, (5), corresponding to static, i.e., non-moving/zero-velocity, structure 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.

2.2.2 Efficient matching

For efficient search, one could resort to: (i) spatiotemporal coarse-to-fine search using spacetime pyramids [23], (ii) evaluation of the template on a coarser sampling of positions in the search volume, (iii) evaluation of a subset of distributions in the template and (iv) early termination of match computation. A drawback of these optimizations 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:

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

with $\star$ denoting cross-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 [12], where the expensive correlation operations in spacetime are exchanged for relatively inexpensive pointwise multiplications in the frequency domain:

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

with $\mathcal{F}\{\cdot\}$ and $\mathcal{F}^{-1}\{\cdot\}$ denoting the Fourier transform and its inverse, resp., 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 $W_{\{\mathbf{T},\mathbf{S}\}}$, $H_{\{\mathbf{T},\mathbf{S}\}}$, $D_{\{\mathbf{T},\mathbf{S}\}}$ 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 [12]. The computational complexity of the frequency domain-based correlation is $O[B W_{\mathbf{S}} H_{\mathbf{S}} D_{\mathbf{S}} (\log_2 D_{\mathbf{S}} + \log_2 W_{\mathbf{S}} + \log_2 H_{\mathbf{S}})]$.

In practice, the overall runtime to compute the entire match 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 26 minutes when computed strictly in the spacetime domain, (10), and 20 seconds (i.e., 10 frames/sec) when computed using the frequency-based scheme, (11), with the the computation of the representation (Sec. 2.1) taking up 16 seconds of the total time. 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, [23] report that their 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 off-line. Also, since the representation construction and matching are highly parallelizable, real to near-real-time performance is anticipated through the use of widely available hardware and instruction sets, e.g., multicore CPUs, GPUs and SIMD instruction sets.

To recapitulate, the proposed approach is given in

Algorithm 1: Action spotting.

Input: T: Query video, S: Search video, τ: Similarity threshold

Output: M: Similarity volume, d: Set of bounding volumes of detected action

`Step 1: Compute spacetime oriented energy representation (Sec. 2.1)`

1. Initialize 3D G_3 steerable basis.
2. Compute normalized spacetime oriented energies for T and S, Eq. (1) - (6).

`Step 2: Spacetime template matching (Sec. 2.2, 2.2.1 and 2.2.2)`

3. (Optional) Weight template representation.
4. Compute M between T and S (Eq. 11).

`Step 3: Find similarity peaks (Sec. 2.2)`

5. Find global maximum in M.
6. Suppress values around the maximum (set to zero).
7. Repeat 5. and 6. until remaining match scores are below τ.
8. Centre bounding boxes, d, with size of template at identified maxima.

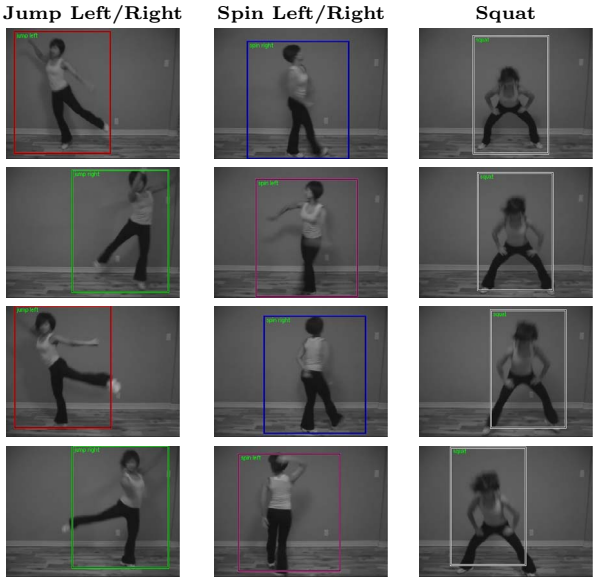

Figure 2. Aerobics routine. The instructor performs four cycles of a dance routine composed of directional jumping, spinning and squatting; each row corresponds to a cycle.

algorithmic terms in Algorithm 1.

3. Empirical evaluation

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 action. Unless otherwise stated, the templates are weighted by $1-(\hat{E}_\epsilon+\hat{E}_{\text{static}})$, see (9), to emphasize the dynamics of the action pattern over the background portion of the template. The constant ϵ is empirically set to 500 for all experiments. All video results and additional examples are available at: `www.cse.yorku.ca/vision/research/action-spotting`.

Figure 2 shows results of the proposed approach on an aerobics routine consisting of 806 frames and a resolution of 361×241 (courtesy of `www.fitmoves.com`). The instructor performs four cycles of the routine composed of directional jumping, spinning and squatting. The first cycle consists of jumping left, spinning right and squatting while the instructor is moving to her left, then a mirrored version while moving to her right, then again moving to her left and concludes moving to her right. Challenging aspects of this example include, fast (realistic) human movements and the instructor introducing nuances into each performance of the same action; thus, no two examples of the same action are exactly alike. The templates are all based on the first cycle of the routine. The jumping right and spinning left templates are defined by mirrored versions of the jumping left and spinning right, resp. Each template is matched with the search target, yielding five similarity volumes in total. Matches in each volume are combined to yield the final detections. All actions are correctly spotted, with no false positives.

Figure 3. Multiple actions in the outdoors. (top, middle) Sample frames for two query templates depicting walking left and two-handed wave actions. A third template for a walking right action is derived from mirroring the walking left template. (bottom) Several example detection results from an outdoor scene containing the query actions.

Figure 3 shows results of the proposed approach on an outdoor scene consisting of three actions, namely, walking left, walking right and two-handed wave. In addition, there are several instances of distracting background clutter, most notably the water flowing from the fountain. The video consists of 672 frames with a resolution of 321×185. The walk right template is a mirrored version of the walk left template. Similar to the previous example, the matches for each template are combined to yield the final detections. All actions are spotted correctly, save one walking left action that differs significantly in spatial scale by a factor of 2.5 from the template; there are no false positives.

Figure 4 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.

Figure 4. Foreground clutter. (top) Sample frames for a one-handed wave query template. Templates for other actions in this figure (two-handed wave, walking left) are shown in Fig. 3. (middle) 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) Sample two-handed wave detection results as action is performed beside and behind chain-linked fence. Foreground clutter takes the form of superimposed static structure when action is behind fence.

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., pseudo-transparency) caused by the chain-linked fence and the query action of two-handed wave. The first and second examples contain 365 and 699 frames, resp., with the same spatial resolution of 321×185 pixels. All actions are spotted correctly; there are no false positives.

For the purpose of quantitatively evaluating the proposed approach, action spotting performance was tested on the publicly available CMU action data set [15]. The data set is comprised of five action categories, namely "pick-up", "jumping jacks", "push elevator button", "one-handed wave" and "two-handed wave". The total data set consists of 20 minutes of video containing 109 actions of interest with 14 to 34 testing instances per category performed by three to six subjects. The videos are 160×120 pixels in resolution. In contrast to the widely used KTH [22] and Weizmann [11] data sets which contain relatively uniform intensity backgrounds, the CMU data set was captured with a handheld camera in crowded environments with moving people and cars in the background. There are large variations in the performance of the target actions, including their distance with respect to the camera.

Results are compared with ground truth labels included with the CMU data set. 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 between 0.6 to 0.97: $Precision = TP/(TP + FP)$ and $Recall = TP/nP$, 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 data set. In evaluation, the same testing protocol as in [15] is used. A detected action is considered a true positive if it has a (spacetime) volumetric overlap greater than 50% with a ground truth label. The same action templates from [15] are used, including the provided spacetime binary masks to emphasize the action over the background.

Figure 5 shows P-R curves for each action with representative action spotting results. The blue curves correspond to the proposed approach, while the red and green curves are results from two baseline approaches (as reported in [15]), namely parts-based shape plus flow [15], and holistic flow [23], resp. In general, the proposed approach achieves significantly superior performance over both baselines, except in the case of two-handed wave. Two-handed wave is primarily confused with one-handed wave, resulting in a higher false positive rate and thus lower precision; nevertheless, the proposed approach still outperforms holistic flow over most of the P-R plot for this action.

4. Discussion and summary

The main contribution of this paper is the representation of visual spacetime via spatiotemporal orientation distributions for the purpose of action spotting. 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 approach is 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 action is depicted in natural conditions. Large geometric spatial (e.g., scale) and temporal (e.g., action execution speed) deformations are not currently handled but may be addressed partly through the use of a multi-scale (spacetime) framework. Additional sources of natural variability include deviations due to local differences in action execution (e.g., performance nuances) and anthropomorphic attributes (e.g., height and shape). Response to these variations motivates future investigation of deformable action templates (e.g., parts-based, cf. [15]), which will allow for flexibility in matching template (sub)components. Also, false detections may be reduced through the integration of complementary cues, e.g., shape. Along these lines, it is interesting to note that the presented empirical results attest that the 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

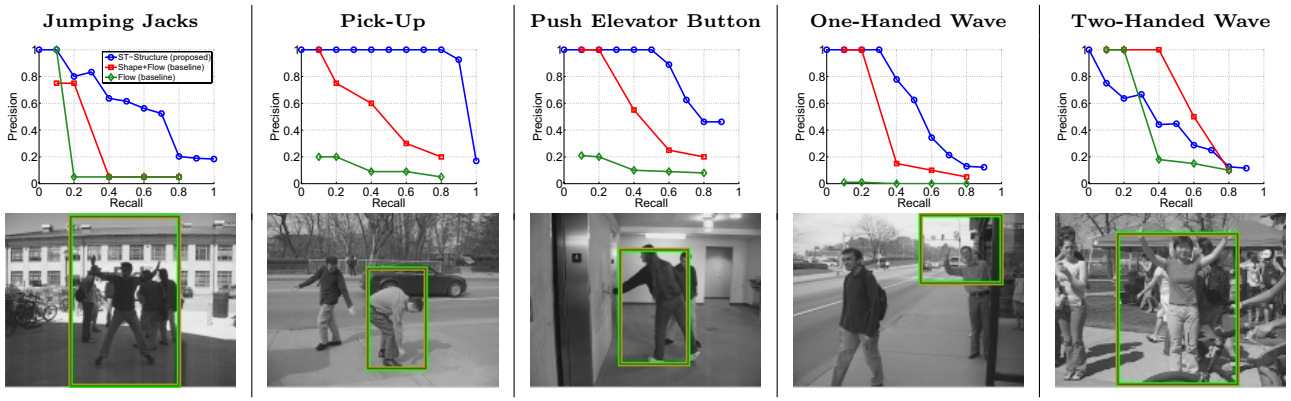

Figure 5. Precision-Recall curves for CMU action data set. (top) Precision-Recall plots; blue curves correspond to the proposed approach, and red and green to the baselines proposed in [15] (shape+flow parts-based) and [23] (holistic flow), resp., as reported in [15]. (bottom) Corresponding example action spotting results recovered by the proposed approach.

as well as individual performance nuances). Such robustness owes 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 an efficient approach to action spotting using a single query template; there is no need for extensive training. The approach is founded on a distributed characterization of visual spacetime in terms of 3D, (x, y, t), spatiotemporal orientation that captures underlying pattern dynamics. Empirical evaluation on a broad set of image sequences, including a quantitative comparison with two baseline approaches on a challenging public data set, demonstrates the potential of the proposed approach.

References

[1] S. Ali, A. Basharat, and M. Shah. Chaotic invariants for human action recognition. In *ICCV*, 2007.

[2] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distribution. *Bull. Cal. Math. Soc.*, 35:99–110, 1943.

[3] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *PAMI*, 23(3):257–267, 2001.

[4] K. Cannons and R. Wildes. Spatiotemporal oriented energy features for visual tracking. In *ACCV*, pages I: 532–543, 2007.

[5] O. Chomat, J. Martin, and J. Crowley. A probabilistic sensor for the perception and the recognition of activities. In *ECCV*, pages I: 487–503, 2000.

[6] K. Derpanis and J. Gryn. Three-dimensional nth derivative of Gaussian separable steerable filters. In *ICIP*, pages III: 553–556, 2005.

[7] K. Derpanis and R. Wildes. Early spatiotemporal grouping with a distributed oriented energy representation. In *CVPR*, 2009.

[8] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *PETS*, pages 65–72, 2005.

[9] A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *ICCV*, pages 726–733, 2003.

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

[11] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *PAMI*, 29(12):2247–2253, 2007.

[12] B. Jähne. *Digital Image Processing*. Springer, 2005.

[13] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. In *ICCV*, 2007.

[14] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *ICCV*, pages I: 166–173, 2005.

[15] Y. Ke, R. Sukthankar, and M. Hebert. Event detection in crowded videos. In *ICCV*, 2007.

[16] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos 'in the wild'. In *CVPR*, 2009.

[17] J. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *IJCV*, 79(3):299–318, 2008.

[18] H. Ning, T. Han, D. Walther, M. Liu, and T. Huang. Hierarchical space-time model enabling efficient search for human actions. *T-CirSys*, 2009.

[19] D. Ramanan, D. Forsyth. Automatic annotation of everyday movements. In *NIPS*, 2003.

[20] M. Rodriguez, J. Ahmed, and M. Shah. Action MACH a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, 2008.

[21] Y. Rubner, J. Puzicha, C. Tomasi, and J. Buhmann. Empirical evaluation of dissimilarity measures for color and texture. *CVIU*, 84(1):25–43, 2001.

[22] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions. In *ICPR*, pages III: 32–36, 2004.

[23] E. Shechtman and M. Irani. Space-time behavior-based correlation - OR - How to tell if two underlying motion fields are similar without computing them? *PAMI*, 29(11):2045–2056, 2007.

[24] E. Simoncelli. *Distributed Analysis and Representation of Visual Motion*. PhD thesis, MIT, 1993.

[25] M. Sizintsev and R. Wildes. Spatiotemporal stereo via spatiotemporal quadric element (stequel) matching. In *CVPR*, 2009.

[26] P. Turaga, R. Chellappa, V. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *T-CirSys*, 18(11):1473–1488, 2008.

[27] R. Wildes and J. Bergen. Qualitative spatiotemporal analysis using an oriented energy representation. In *ECCV*, pages II: 768–784, 2000.

[28] Y. Yacoob and M. Black. Parameterized modeling and recognition of activities. *CVIU*, 73(2):232–247, 1999.