Nested Motion Descriptors

Jeffrey Byrne
University of Pennsylvania, GRASP Lab
Systems and Technology Research
jeffrey.byrne@stresearch.com

Abstract

A nested motion descriptor is a spatiotemporal representation of motion that is invariant to global camera translation, without requiring an explicit estimate of optical flow or camera stabilization. This descriptor is a natural spatiotemporal extension of the nested shape descriptor [2] to the representation of motion. We demonstrate that the quadrature steerable pyramid can be used to pool phase*, and that pooling phase rather than magnitude provides an estimate of camera motion. This motion can be removed using the log-spiral normalization as introduced in the nested shape descriptor. Furthermore, this structure enables an elegant visualization of salient motion using the reconstruction properties of the steerable pyramid. We compare our descriptor to local motion descriptors, HOG-3D and HOG-HOF, and show improvements on three activity recognition datasets.*

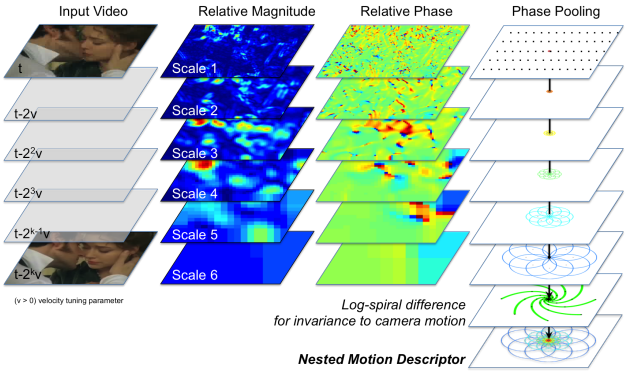

Figure 1. Nested Motion Descriptors (NMD). (left) Compute relative magnitude and phase for orientations and scales for a set of frames, (right) Pool the robust component velocity derived from relative phase in a set of circular pooling regions all intersecting at the center interest point. Log-spiral normalization computes the difference between phases in neighboring scales and positions along a log-spiral curve. The phase pooling aggregates component velocities, so this difference computes an acceleration which represents local motion which is invariant to constant velocity of the camera.

1. Introduction

The problem of activity recognition is a central problem in video understanding. This problem is concerned with detecting actions in a subsequence of images, and assigning this detected activity a unique semantic label. The core problem of activity recognition is concerned with the representation of *motion*, such that the motion representation captures the informative or meaningful properties of the activity, and discards irrelevant motions due to camera or background clutter.

A key challenge of activity recognition is motion representation in *unconstrained video*. Classic activity recognition datasets [21] focused on tens of actions collected with a static camera of actors performing scripted activities, however the state-of-the-art has moved to recognition of hundreds of activities captured with moving cameras of "activities in the wild" [12][19][15]. Moving cameras exhibit unconstrained translation, rotation and zoom, which introduces motion at every pixel in addition to pixel motion due to the foreground activity. The motion due to camera movement is not informative for the activity, and has been shown to strongly affect activity representation performance [8].

Recent work has focused on motion descriptors that are invariant to camera motion [11, 7, 31, 8, 26, 25, 27, 16, 29]. Local spatiotemporal descriptors such as, such as HOG-HOF [3, 14] or HOG-3D [10], have shown to be a useful motion representation for activity recognition. However, these local descriptors are not invariant to dominant camera motion. Recent work has focused on aggregating these local motion descriptors into *dense trajectories*, where optical flow techniques are used to provide local tracking of each pixel. Then, the local motion descriptors are constructed using differences in the flow field, and then are concatenated along a trajectory for invariance to global motion. However, these approaches all rely on estimation of the motion field using optical flow techniques, which have shown to introduce artifacts into a video stream due to an early commitment to motion or over-regularization of the motion field,

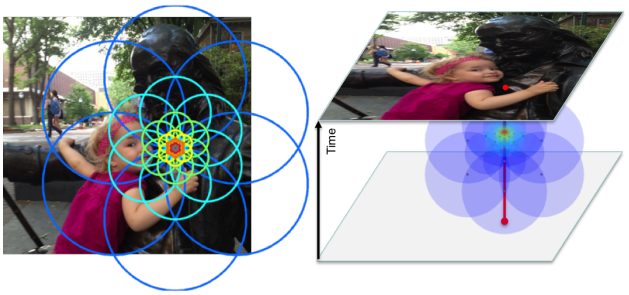

Figure 2. From nested shape descriptors to nested motion descriptors. Nested shape descriptors pool oriented and scaled gradients magnitude which captures the contrast of an edge in an image. Nested motion descriptors pool *relative phase* which captures *translation* of an edge. Projecting the spatiotemporal structure of the nested motion descriptor onto a single image will form the structure of the nested shape descriptor.

which can corrupts the motion representation.

In this paper, we propose a new family of binary local motion descriptors called *nested motion descriptors*. This descriptor provides a representation of *salient motion* that is invariant to global camera motion, without requiring an explicit optical flow estimate. The key new idea underlying this descriptor is that appropriate sampling of scaled and oriented gradients in the complex steerable pyramid exhibits a *phase shift* due to camera motion. This phase shift can be removed by a technique called a *log-spiral normalization*, which computes a phase difference in neighboring scales and positions, resulting in a relative phase where the absolute global image motion has been removed. This approach is inspired by phase constancy [5], component velocity [4] and motion without movement [6, 24], which uses phase shifts as a correction for translation without an explicit motion field estimate.

2. Related Work

The literature on motion representation can be decomposed into approaches focused on local motion descriptors, mid-level motion descriptors or global activity descriptors. In this section, we will focus on local motion representations only, which are most relevant to this paper.

A *local motion descriptor* is a representation of the local movement in a scene centered at a single interest point in a video. Examples of local motion descriptors include HOG-HOF [3, 14], cuboid [17], extended SURF [30] and HOG-3D [10]. These descriptors construct spatiotemporal oriented gradient histograms over small spatial and temporal support, typically limited to tens of pixels spatially, and a few frames temporally. HOG-HOF includes a histogram of optical flow [3, 14], computed over a similar sized spatiotemporal support. Furthermore, recent evaluations have shown that activity recognition performance is significantly improved by considering dense regular sampling of descriptors [28][1], rather than sparse extraction at interest points.

An interesting recent development has been the development of local motion descriptors that are invariant to dominant camera motion. A translating, rotating or zooming camera introduces global pixel motion that is irrelevant to the motion of the foreground object. Research has observed that this camera motion introduces a global translation, divergence or curl into the optical flow field [8], and removing the effect of this global motion significantly improves the representation of foreground motion for activity recognition. The motion boundary histogram [3, 26, 25] computes a global motion field from optical flow, then computes local histograms of derivatives of the flow field. This representation is sensitive to local changes in the flow field, and insensitive to global flow. Motion interchange patterns [11, 7, 31] compute a patch based local correspondence to recover the motion of a pixel, followed by a trinary representation of the relative motion of neighboring patches. Finally, dense trajectories [26, 25, 27] concatenate HOG-HOF and motion boundary histograms for a tracked sequence of interest points forming a long term trajectory descriptor. The improved dense trajectories [27] with fisher vector encoding is the current state-of-the-art on large datasets for action recognition [29].

3. Nested Motion Descriptors

A nested motion descriptor is a representation of salient motion in a video that is invariant to camera motion. The nested motion descriptor is an extension of the nested shape descriptor [2] to the representation of motion. Figure 2 shows that while the nested shape descriptor pools the magnitude of edges, the nested motion descriptor pools phase gradients which captures translation of edges in a video. In this section, we describe this construction in detail.

3.1. Overview

Figure 1 provides an overview of the construction of the nested motion descriptor. This procedure is summarized as a three step process: bandpass filtering, spatiotemporal phase pooling and log-spiral normalization. First, *bandpass filtering* is performed to decompose each image in a video into a set of orientation and scale selective subbands using the complex steerable pyramid [23, 22, 18]. The complex steerable pyramid includes basis filters in quadrature pairs, which allows for magnitude and phase estimation for each subband. We compute the relative magnitude and relative phase for each subband from a current frame to a past frame. This relative bandpass response is visualized in figure 1. We compute relative magnitude and phase for scales following a log scale, so that we compute a large scale bandpass response for frames further away in time. This encodes a fixed velocity tuning for a velocity parameter ν. The com-

plex steerable pyramid decomposition is described in detail in the supplementary material.

Relative magnitude and phase provide measurements of *speed and direction of motion* in a video. An example is shown in figure 1 (middle) of "kiss" from the human motion database [12]. In this example, the man on the left tilts his head and moves in towards the woman on the right. Observe that there is small scale motion of the man's sideburns and ear, medium scale motion of the collar and woman's eyes, and large scale motion of the two heads moving towards each other. The relative magnitude over various scales captures this motion. Similarly, the relative phase encodes a spatial translation from frame t to $t-k$. The relative phase is shown on the scale $[-\pi, \pi]$ where zero phase is green, negative phase is blue and positive phase is red. The phase of the mid and large scale motions encode the movement of the faces. Furthermore, at the largest scale, observe that there are two motions present, of the two heads moving towards each other.

Second, we perform *phase pooling*. We derive the relationship between phase gradients and component velocity, such that pooling component velocity is equivalent to pooling phase gradients. Furthermore, we derive a robust form of the component velocity using phase stability, to provide robust measurements of component velocities in regions of unstable phase. We define a set of pooling regions to pool the component velocity in neighboring spatial and temporal regions, to provide invariance to local geometric transformations. Each of the pooling regions is centered at an interest point, and the pooling regions are uniformly distributed in angle around the interest point. Each pooling region is represented by a single component velocity, and all orientations and scales are concatenated into a single nested motion descriptor for the interest point. This is visualized in figure 1 by the "collapsing" of the descriptor across scales into a combined descriptor at the bottom of the figure. This pooling and sampling of subband component velocity is the primary construction of the nested motion descriptor.

Third, we perform *log-spiral normalization*. Relative phase or *phase gradients* are proportional to the motion in an image. This motion could be due to the salient motion of a foreground object, or due to the global motion of the camera. Observe that the global motion of the camera introduces pixel motion that is a composition of global translation, rotation and scale. In these cases, the motion field in a local patch is uniformly offset, so that all vectors in the motion field in this patch are offset by a fixed bias due to the camera motion. The relative phase is also offset by a fixed constant. We can remove this constant by computing a phase difference with neighbors in position and scale. This is the goal of the log-spiral normalization, which computes a phase difference to remove this fixed bias due to camera motion. The log-spiral normalization procedure is outlined in figure 1 (bottom right), with the spiral like arrangement showing the differences to be computed along this spiral.

In this section, we describe each of these stages of processing in more detail. The supplementary material provides additional background material on complex steerable pyramid and the relationship between component velocities and phase gradients. The reader is referred to this material for additional detail.

3.2. Phase Gradients and Component Velocity

The complex steerable pyramid [23, 22, 18] is an overcomplete decomposition of an image into orientation and scale selective subbands. The orientation subbands exhibit a steerability property such that the response to an arbitrary orientation is a linear combination of basis subbands. Furthermore, a complex steerable pyramid includes basis filters in quadrature pairs, such that each basis filter is further decomposed into a an oriented filter and it's Hilbert transform, forming an in-phase and quadrature component shifted by 90^o in phase. From this decomposition, it is straightforward to compute a phase and magnitude response at many scale and orientation selective subbands. In this section, we show the relationship between phase and velocity.

This relationship between phase and motion has been used in phase based optical flow methods [4, 5] to enforce the *phase constancy* constraint [4], such that feasible optical flow solutions are constraint to lie on contours of constant phase. This constraint has shown to be more stable than the more common brightness constancy constraint [9, 5] over ranges of shape deformation and lighting. The phase constancy constraint is given by

$$\bigtriangledown \phi(x,t) \bullet \vec{v} = 0 \tag{1}$$

where $\bigtriangledown\phi(x,t) = [\frac{\partial\phi}{dx},\ \frac{\partial\phi}{dy},\ \frac{\partial\phi}{dt}]^T$ is the *phase gradient* and $\vec{v} = [\frac{\partial x_0}{dt},\ \frac{\partial y_0}{dt},\ 1]^T$ is the *component velocity* at point (x_0, y_0). Rearranging terms

$$\frac{\partial\phi}{dx}v_x + \frac{\partial\phi}{dy}v_y = -\frac{\partial\phi}{dt} \tag{2}$$

where we use the shorthand notation $\vec{v} = [v_x,\ v_y,\ 1]$ for the partial derivatives of component velocity and similarly $\bigtriangledown\phi(x,t) = [\phi_x,\ \phi_y,\ \phi_t]^T$ for the phase gradient. The phase constancy constraint states that the projection of the component velocity onto the spatial phase gradient is equal to the negative temporal phase gradient. This is identical to the classic brightness constancy constraint, using local phase instead of local brightness. The phase constancy constraint in (2) shows the explicit relationship between the phase gradient and velocity.

This method can be used to estimate the component velocity for each tuned orientation and scale $B_{\omega,\theta}$. We use the notation $\vec{\phi} = [\phi_x,\ \phi_y]^T$ to denote the spatial phase gradient, then the spatial phase gradient defines a unit vector

$\hat{n} = [\frac{\phi_x}{|\vec{\phi}|}, \frac{\phi_y}{|\vec{\phi}|}]^T$. The unit vector constraints the direction of the component velocity, due to the dot product in the phase constancy constraint. The velocity magnitude α can be determined directly from (2):

$$\alpha = \frac{-\phi_t}{|\vec{\phi}|} \tag{3}$$

where $\vec{\phi} = [\phi_x,\ \phi_y]$ is the spatial phase gradient. This is a single equation in a single unknown for the velocity scale α. Given the observed phase gradient, the component velocity is estimated $\vec{v} = \alpha\hat{n}$.

The component velocity (3) is a function of only phase gradients which can be computed efficiently from the complex steerable pyramid. The bandpass response in the complex steerable pyramid for a given tuned orientation and scale at time t is denoted $B^t_{\omega,\theta}$. To simplify notation, when the bandpass orientation and scale (ω, θ) is implied, let this bandpass response be written as $B^t_{\omega,\theta} = B_t$. The phase gradient is given by

$$\bigtriangledown \phi = \frac{Im(B^* \Delta B)}{|B|^2} \tag{4}$$

where $Im(z)$ is the imaginary component of the complex number z, and B^* is the complex conjugate of the complex valued bandpass response [4].

3.3. Robust Component Velocity

It is important to discuss the *stability* of phase based component velocity estimation. Fleet and Jepson [9, 5] suggest a threshold on a function of the magnitude response to discard regions with poor phase stability. They show that a sufficient statistic for a robust phase estimate is the ratio between the spatial derivative of magnitude and the absolute magnitude. In other words, we require a small change in magnitude relative to the absolute magnitude in order to have stable phase estimate.

$$P = \{q \mid \frac{|\rho_x(q)|}{\rho(q)} < \tau,\ q \in I\} \tag{5}$$

The set P is a set of interest points in an image I such that each interest point satisfies the constraint for phase stability. A feasible interest point is one that has a small spatial change in magnitude (e.g. a local maxima of magnitude, at the phase zero crossing) and has a large edge magnitude. This constraint discards regions of low contrast (small denominator) and non-maximum edges (large numerator), leaving interest points that have sufficiently stable phase characteristics for computing component velocity.

The stability constraint in (5) be combined with the phase gradient (4) into a single measurement of *robust phase gradient* $\bigtriangledown\hat{\phi}$

$$f(\rho, \tau, \beta) = \frac{1}{1 + \exp(-\beta(\tau - \frac{|\rho_x|}{\rho}))} \tag{6}$$

$$\bigtriangledown\hat{\phi} = f(\rho, \tau, \beta) \bigtriangledown \phi = \frac{\bigtriangledown\phi}{1 + \exp(-\beta(\tau - \frac{|\rho_x|}{\rho}))} \tag{7}$$

The logistic function in (6) provides a soft threshold for the stability constraint. The robust phase gradient is equal to $\bigtriangledown\phi$ when $\frac{|\rho_x|}{\rho} \ll \tau$, and smoothly transitions to zero as $\frac{|\rho_x|}{\rho} \gg \tau$. The parameter β encodes the sharpness of the transition of the logistic function from zero to one.

This estimate of robust phase gradient can be used to define a *robust component velocity*. Following the definition of component velocity in (3), and replacing the phase gradients with the robust phase gradients in (7), we define the robust component velocity as

$$\hat{\alpha} = \frac{-\hat{\phi}_t}{|\vec{\hat{\phi}}|} \tag{8}$$

Intuitively, this function provides a measurement of component velocity that is equal to the observed velocity if the magnitude is sufficient. However, if the magnitude is not sufficient and the phase is unstable, such as a region of low contrast, then the function will provide a measurement of zero velocity. This formulation of robust component velocity is a new contribution of this work.

Figure 3 shows and example of the phase stability and robust phase gradient. In this example, a golfer is in the middle of the backswing and the camera is panning from left to right to begin following the ball. We show the magnitude and phase for an oriented bandpass response tuned to two octave scales and 0^o orientation. The observed temporal phase gradient is very noisy due to regions of poor stability where the magnitude is small or non-maximum. The phase stability in (5) can be used to identify the regions in the imagery with stable phase, which is shown in the grayscale image such that white pixels are stable, and black are unstable. Finally, the robust phase gradient is computed using this stability constraint as in (7) resulting in stable phase measurements. The figure shows that the stable phase gradient is much less noisy and clearly reflects the true motion of the background and the swing of the golfer.

3.4. Robust Phase Pooling

Spatiotemporal phase pooling refers to the aggregation or accumulation of phase gradients over neighboring positions and times. The pooling regions over which the accumulation occurs are represented as spheres in a 3D spatiotemporal volumes (x, y, t) where (x, y) are spatial image support in pixels and t is the temporal support in frames of a video. The radius of the sphere defines the spatial and

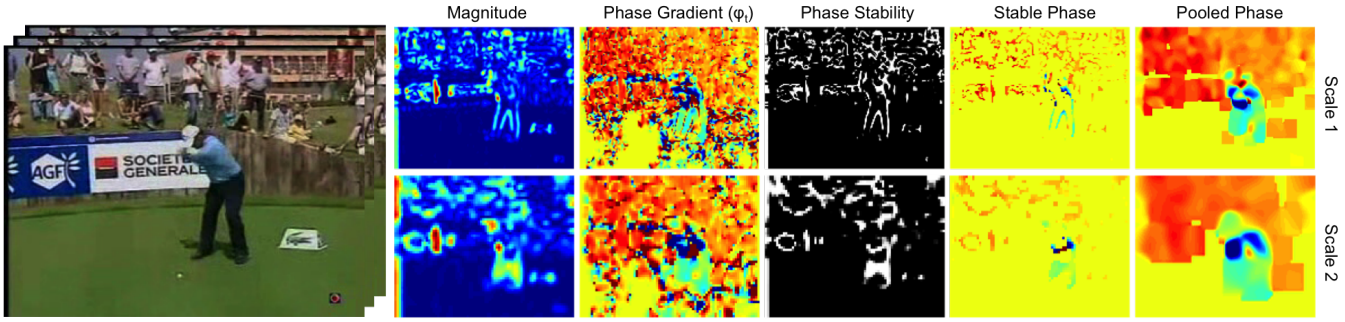

Figure 3. Robust Phase Pooling. The temporal phase gradient is noisy due to the measurement of phase in regions where phase is unstable, such as the region on the grass and in the crowd. The phase stability measure provides an estimate of locations of stable phase. Only the stable phase is used for pooling, resulting in pooled phase that captures the motion of the background and foreground of the golfer in the scene. This pooled phase is used to construct the nested motion descriptor.

temporal support of the aggregation. Figure 2 shows perspective views of the spatiotemporal pooling regions for the nested motion descriptor.

Spatiotemporal pooling in the nested motion descriptor is constrained such that the temporal projection of pooling regions is equivalent to the nested shape descriptor [2]. Formally, the spatiotemporal nested pooling is defined as follows. We will use the notation and conventions defined in [2], where sets of spheres are grouped into lobes forming an Hawaiian earring when projected onto the (x,y) plane. The descriptor exhibits n-fold rotational symmetry so that there are n lobes equally spaced in angle. The notation $\mathbb{K}_n(i,j)$ refers to the sphere in the i^{th} lobe at j^{th} scale, with center $c_{ij} = [2^j cos(i\frac{2\pi}{n}),\ 2^j sin(i\frac{2\pi}{n}), 2^j\nu]^T$ and radius $r_{ij} = [2^i,\ 2^i,\ 2^i\nu]^T$ in (x,y,t) spatiotemporal volume. The parameter ν is the velocity tuning of the NMD, which "squashes" the descriptor temporally to tune to faster or slower motion.

Finally, we perform pooling of robust phase gradients within these spherical pooling regions. Recall that the definition of the robust phase gradients uses the fact that some regions of the image are unstable, and do not provide reliable phase estimates. So, phase cannot just be accumulated over each pooling region, as there may be different number of stable phase estimates in each region. To compensate, we pool robust phase gradients, but normalize by the total phase stability measure in the pooling region. This phase pooling is equivalent to the mean robust phase gradient within the pooling region. Figure 3 shows an example of this pooling in the final column. This phase pooling is used to construct the robust component velocity and the nested motion descriptor.

3.5. Construction of the Nested Motion Descriptor

Finally, we can pull together the results from the previous sections to construct a nested motion descriptor at an interest point as follows. Let $B^t_{\omega,\theta}$ be a bandpass response at scale ω and orientation θ at time t, for each frame in a video clip as computed from the complex steerable pyramid. Next, compute the phase gradients for each bandpass response following (4), and compute the robust phase gradient following (7). This stable phase is pooled using the spatiotemporal pooling in section 3.4 for a given spatiotemporal pooling support $\mathbb{K}_n$. Finally, the robust component velocity is computed as in (8) for the pooled phase gradients. Let the robust component velocity be indexed $\hat{\alpha}^t_{ij}(q)$ for orientation i and scale j at pixel q, where the phase gradient is computed using the current frame and frame t. Then, the nested motion descriptor is constructed from pooled robust component velocities, normalized by the stability constraint:

$$d(i,j,k,t) = \frac{\Sigma_{q\in\mathbb{K}_n(j,k)}\ \hat{\alpha}^t_{ik}(q)}{\Sigma_{q\in\mathbb{K}_n(j,k)}\ f(q)} \tag{9}$$

$$\hat{d}(i,j,k) = d(i,j,k,t-2^k\nu) - d(i,j{-}1,k{-}1,t-2^k\nu) \tag{10}$$

$$D(i,j,k) = \begin{cases} 1 & \text{if } \hat{d}(i,j,k) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

Equation (9) is *robust component velocity pooling*. The descriptor $d(i,j,k,t)$ is the pooled component velocity for orientation subband i, lobe j and lobe scale k at frame t. Observe that the bandpass scale k is equal to the pooling support radius k. In other words, support regions with radius 2^k pool orientation subbands over octave scales k, so we pool coarser gradients over larger supports. Furthermore, the normalization constant is the pooled phase stability constraint in (6). This provides a weighted mean component velocity within the pooling region, where the weight is provided by the phase stability.

Equation (10) is *logarithmic spiral normalization*. This log-spiral normalization computes the difference between component velocities at neighboring scales and positions within the same frame. Observe that there is a coupling between the frame offset, pooling scale and bandpass scale,

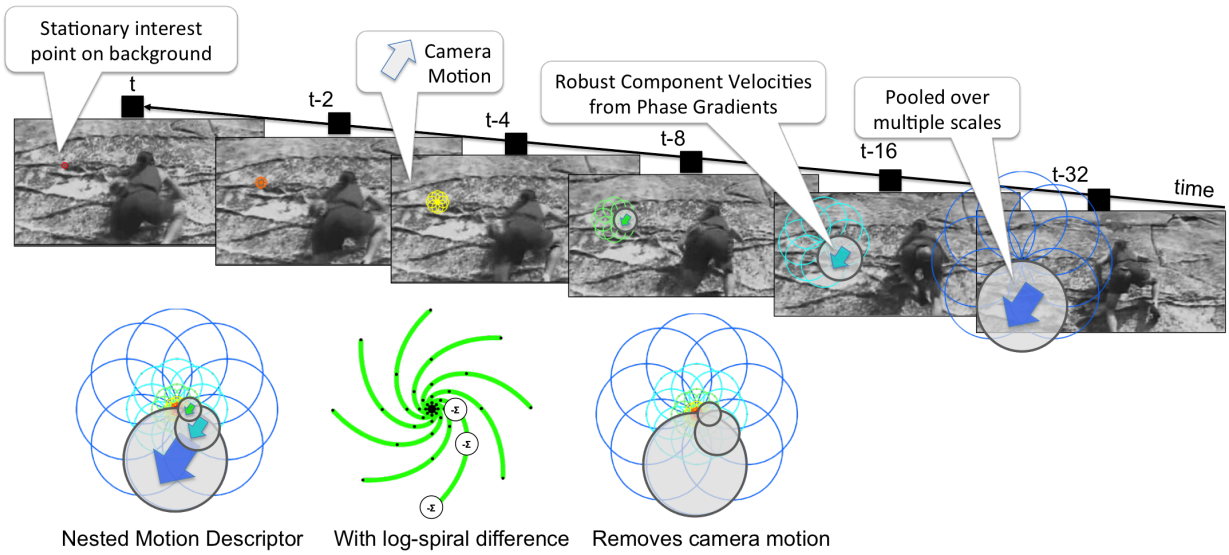

Figure 4. The nested motion descriptor is invariant to global camera motion. (top) A video sequence of a rock climber where the camera is following the climber up the rock face. For a given fixed interest point on the background, we compute the nested motion descriptor. Observe that the robust component velocities for this interest points are the same. (bottom) When computing the log-spiral difference, the constant velocity due to the camera motion is removed, leaving only *acceleration*.

since all depend on k. This results in pooling coarser velocities over larger supports. We discuss in the next section how this normalization provides invariance to camera motion. This log-spiral normalization is also discussed in more detail in the supplementary material.

Finally, equation (11) is *binarization*. A nested motion descriptor can be binarized by computing the sign of (10). This constructs a nested motion descriptor with binary entries. This is an optional step which can be used to provide compact representation.

The final nested motion descriptor D from (11) is a binary vector of length $(R \times |\mathbb{K}| \times |K|)$ for R orientation bands over $|\mathbb{K}|$ lobes and $|K|$ supports per lobe. For example, for eight orientation subbands, five nested supports, and six lobes has dimensionality $(8 \times 6 \times 5) = 240$. The nested motion descriptor can also be real valued using (10), without the final binarization step.

3.6. Invariance to Camera Motion

In this section, we describe how the log-spiral normalization of the nested motion descriptor provides invariance to global camera motion. The key intuition for this procedure is that each dimension of the NMD encodes the robust component velocity of estimated at a specific orientation and scale. The log-spiral normalization computes a difference between neighboring scales and positions in the NMD, within the same frame. If both of these dimensions are moving with the same velocity, due to the global camera motion, then the difference will remove this effect. Basically, the log-spiral difference is computing an *local acceleration* or second order derivative between neighboring velocities pooled in the nested motion descriptor. Acceleration is invariant to constant velocity, so if the camera is translating with a constant velocity, the descriptor will be invariant to this motion.

Figure 4 shows an example of the invariance to the dominant camera motion. This figure shows a video sequence of a rock climber where the camera is following the climber up the rock face. This introduces constant velocity motion in the background due to the camera motion. We show a single interest point on the background to show that this effect of the motion from the camera is removed. We compute the robust component velocities using the nested motion descriptor construction in the previous section. Observe that each pooling region on this background interest point result in the same component velocity. This is the same due to the global motion of the camera. When we compute the log-spiral difference, this constant velocity is removed, resulting in robust component velocities of zero.

3.7. Motion Visualization

We can visualize motion representation of the NMD as a saliency map using the steerable pyramid reconstruction. A saliency map is a real valued scalar field that encodes the salience of regions in an image or video. The nested motion descriptor can be used to compute a saliency map in a very simple manner. Recall that the nested motion descriptor requires the construction of a quadrature steerable pyramid to compute multiscale oriented gradients. Given this pyramid, replace the orientation and scale bands with

the clipped mean square response of the NMD for each orientation and lobe. Then, replace the low pass response of the steerable pyramid with the squared Laplacian filter response, to implement a center surround difference. Finally, reconstruct the image from this saliency pyramid. In short, a motion saliency map is the image reconstructed from the squared response of the nested motion descriptor.

Formally, let a steerable pyramid $B = \{I_0, B_{ij}\ ; i \leq R, j \leq S\}$ for orientation bands B_{ij} over R orientations i and S scales j and lowpass residual image I_0. Each band B_{ij} encodes the oriented gradient response at orientation i and scale j. Furthermore, let $\hat{d}$ be a log-spiral normalized nested motion descriptor constructed following eq. 9 and 10, computed densely at each pixel. Then, let

$$\hat{B}_{ij} = max(\sum_j \hat{d}(i,j,k)^2,\ \tau) \tag{12}$$

$$\hat{I}_0 = (I_0 * L)^2 \tag{13}$$

where L is a 3x3 Laplacian kernel, $*$ is the convolution operation, and τ is a clipping threshold for the maximum squared difference. These are collected as subbands in a steerable pyramid $\hat{B} = \{\hat{I}_0,\ \hat{B}_{ij}\}$, and these bands are used to reconstruct an image using the standard steerable pyramid reconstruction algorithm, where the filters used for reconstruction are the magnitude of the quadrature pair. This reconstructed image is a saliency map. Finally, a saliency video is encoded from the set of saliency maps computed from the video, and rescaled so that the maximum saliency response is encoded as red.

4. Experimental Results

In this section, we show results for applying nested motion descriptor to the task of activity recognition. We focus on three datasets, and compare results for a simple bag-of-words classification framework, to highlight the performance differences due to motion descriptors only.

We compare performance of the nested motion descriptors to HOG-HOF [13] and HOG-3D [10]. The evaluation in [28] showed that HOG-HOF and HOG-3D outperformed cuboid and dense SURF, so we limit our evaluation to these two descriptors. Furthermore, the improved dense trajectories consider HOG-HOF as the local motion descriptor extracted along the trajectory, so we use this as our baseline.

The datasets chosen for this evaluation span the complexity representative of classic and modern activity recognition problems: KTH actions dataset [21] (2004) is a classic dataset, UCF sports actions dataset [20] (2008) has nine activity classes, but these videos are collected in unconstrained television footage, and human motion database (HMDB) [12] (2011) is representative of a modern dataset with over fifty actions in unconstrained video.

Descriptor	KTH Actions	UCF Sports	HMDB
HOG-HOF	0.81	0.62	0.23
HOG-3D	0.86	0.75	0.24
NMD	0.87	0.77	0.25

Table 1. Mean average precision (mAP) results for activity recognition. Results show that the nested motion descriptor (NMD) outperforms the baseline on all classes.

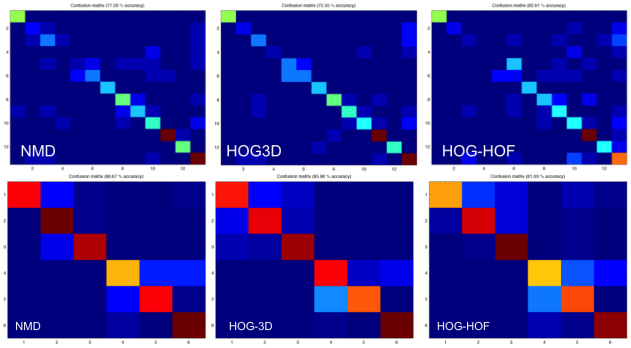

Figure 5. (top) Activity recognition results on UCF sports actions. The class index for each result: 'Diving-Side', 'Golf-Swing-Back','Golf-Swing-Front', 'Golf-Swing-Side', 'Kicking-Front', 'Kicking-Side' , 'Lifting', 'Riding-Horse', 'Run-Side', 'SkateBoarding-Front', 'Swing-Bench', 'Swing-SideAngle' ,'Walk-Front'. (bottom) KTH actions

4.1. Activity Recognition

The overall results are shown in table 1. We report mean classification rate results over all activity classes for activity recognition using the experimental framework is described in detail in the supplementary material.

Results show that the nested motion descriptor (NMD) outperforms the baseline on all datasets. These results are consistent with reported results in the literature using bag-of-words framework, albeit at a lower overall classification rate. We believe this is due to the smaller total vocabulary size (600 vs. 4000 in [28]), however the relative performance change across the dataset is consistent. The best performance is on the KTH actions dataset which does not contain any global camera motion, the second best is on UCF sports which contains camera motion but a limited number of object classes. The worst performance is on unstabilized HMDB, due to the large number of classes. However, we observe that the NMD does still provide improved performance over the baseline descriptors.

Figure 5 shows confusion matrices for UCF sports and KTH actions. Recall that this dataset requires leave one out cross validation results due to the limited number of training examples per class. We observed that this dataset includes a significant background context that affects the results for comparing motion descriptor. Specifically, the "Kicking-Front" and "Kicking-Side" classes contains wide open grass fields with strong field line markers. Observe that the HOG-

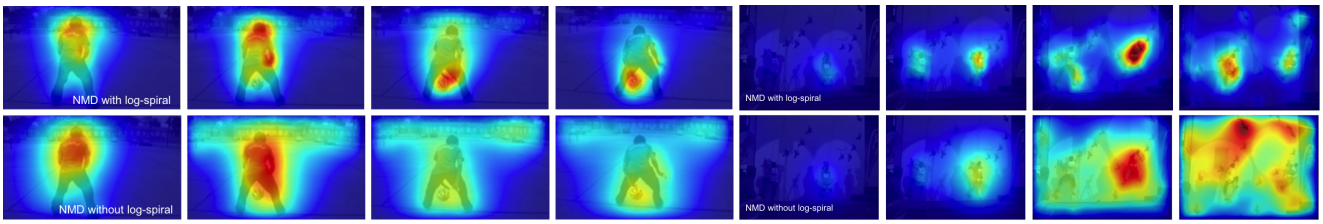

Figure 6. The nested motion descriptor represents salient motion in video. We show a semitransparent saliency map for motion overlayed on each frame of video. This saliency map shows salient responses in red and non-salient in blue. (top left) Salient motion for HMDB "basketball dribbling" using NMD with log spiral normalization. (bottom left) NMD without log-spiral normalization includes motion of the camera. (top right) Motion saliency for HMDB "rock climbing". (bottom right) NMD without log-spiral normalization. The log-spiral normalization suppresses the significant camera motion in the scene focusing on the salient motion of the rock climbers.

3D descriptor confuses only kicking-front and kicking-side, while the NMD performs poorly on this class but better on all other classes. We hypothesize that this is due to the context of the large football fields on which this action takes place, rather than the motion of the foreground itself. The NMD suppresses the motion on the ground due to the dominant camera motion, while the HOG-3D descriptor leverages this context that is unique to these two classes. If we remove these biased classes from the aggregate scores, we see that the NMD outperforms the HOG-3D using motion only on the remaining classes, and these are the score reported. However, this result does highlight the need for a composite descriptor that can leverage features from many different sources, including the surrounding context of the background.

4.2. Motion Visualization Results

In this section, we show qualitative results applying the visualization of salient motion captured by the NMD as described in section 3.7. We show results for a sampling of videos from the KTH actions and HMDB datasets, and compare qualitative results with and without the log-spiral normalization. These results demonstrate the effectiveness of the log-spiral normalization in representation of salient motion and suppressing the effect of camera motion. Additional video results are provided in the supplementary material.

Figure 6 shows an example of four frames of basketball dribbling from HMDB. The top row shows the output of the motion saliency using the NMD, and the bottom row shows the same output using the NMD without the log spiral normalization. This clip contains large scale and small scale motion of the body and hands of the player, as well as global camera motion down and to the left. The colors encode the saliency map such that red is salient and blue is not-salient. Observe that the salient motion extracted using this technique highlight the small motions of dribbling the basketball and not the large motions due to the camera.

Figure 6 (right) shows an example of rock climbing from the HMDB. In this example, two rock climbers are racing to the top of an indoor rock climbing wall and the camera follows the climbers up the wall introducing large camera motion up and to the right. The bottom row shows that without the log-spiral normalization, the background motion tends to dominate the motion representation which manifests as motion everywhere in the scene. The top row shows that the log-spiral normalization is able to suppress this dominant motion so that the motion of the climbers pops out from the background.

5. Summary

In this paper, we introduced the nested motion descriptor for representation of salient motion. We motivated the construction of this descriptor using phase based optical flow, we described the construction of the descriptor and showed that the log-spiral normalization provides invariance to dominant camera motion. Furthermore, we showed visualization of this motion representation for videos with large camera motions, to show that the descriptor is removing the dominant camera motion. Finally, we showed improved performance over the state of the art in local motion descriptors for activity recognition, showing the representational capabilities of this descriptor.

References

[1] P. Bilinski and F. Bremond. Evaluation of local descriptors for action recognition in videos. In *International Conference on Computer Vision Systems*, Sophia Antipolis, France, 2011. 2

[2] J. Byrne and J. Shi. Nested shape descriptors. In *ICCV*, 2013. 1, 2, 5

[3] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *ECCV*, 2006. 1, 2

[4] D. Fleet and A. Jepson. Computation of component image velocity from local phase information. *International Journal of Computer Vision*, 5(1):77–104, 1990. 2, 3, 4

[5] D. Fleet and A. Jepson. Stability of phase information. *IEEE Trans on Pattern Anal. and Mach. Intell. (PAMI)*, 15(12):1253–1268, 1993. 2, 3, 4

[6] W. Freeman, E. Adelson, and D. Heeger. Motion without movement. *ACM Computer Graphics, (SIGGRAPH'91)*, 25(4):27–30, July 1991. 2

[7] Y. Hanani, N. Levy, and L. Wolf. Evaluating new variants of motion interchange patterns. In *CVPR workshop on action similarity in unconstrained video*, 2013. 1, 2

[8] M. Jain, H. Jegou, and P. Bouthemy. Better exploiting motion for better action recognition. In *CVPR*, 2013. 1, 2

[9] A. Jepson and D. Fleet. Phase singularities in scale-space. *Image and Vision Computing Journal*, 9(5):338–343, 1991. 3, 4

[10] A. Klaser, M. Marsza?ek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, 2008. 1, 2, 7

[11] O. Kliper-Gross, Y. Gurovich, T. Hassner, and L. Wolf. Motion interchange patterns for action recognition in unconstrained videos. In *ECCV*, 2012. 1, 2

[12] H. Kuhne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition. In *ICCV*, 2011. 1, 3, 7

[13] I. Laptev. On space-time interest points. *IJCV*, 2005. 7

[14] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008. 1, 2

[15] M. Marszałek, I. Laptev, and C. Schmid. Actions in context. In *IEEE Conference on Computer Vision & Pattern Recognition*, 2009. 1

[16] X. Peng, Y. Qiao, Q. Peng, and X. Qi. Exploring motion boundary based sampling and spatial-temporal context descriptors for action recognition. In *BMVC*, 2013. 1

[17] G. C. Piotr Dollr, Vincent Rabaud and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *ICCV VS-PETS 2005*, 2005. 2

[18] J. Portilla and E. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, 2000. 2, 3

[19] K. K. Reddy and M. Shah. Recognizing 50 human action categories of web videos. *Machine Vision and Applications Journal (MVAP)*, 2012. 1

[20] M. Rodriguez, J. Ahmed, and M. Shah. Action mach: A spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, 2008. 7

[21] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *ICPR*, 2004. 1, 7

[22] E. Simoncelli and W. Freeman. The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *IEEE Second Int'l Conf on Image Processing*, 1995. 2, 3

[23] E. Simoncelli, W. Freeman, E. Adelson, and D. Heeger. Shiftable multi-scale transforms. *IEEE Trans. Info. Theory*, 2(38):587–607, 1992. 2, 3

[24] N. Wadhwa, M. Rubinstein, F. Durand, and W. T. Freeman. Phase-based video motion processing. *ACM Trans. Graph. (Proceedings SIGGRAPH 2013)*, 32(4), 2013. 2

[25] H. Wang, A. Klaeser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *IJCV*, 2013. 1, 2

[26] H. Wang, A. Klser, C. Schmid, and L. Cheng-Lin. Action recognition by dense trajectories. In *CVPR*, 2011. 1, 2

[27] H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV*, 2013. 1, 2

[28] H. Wang, M. M. Ullah, A. Klser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009. 2, 7

[29] H. Weng and C. Schmid. Lear-inria submission for the thumos workshop. In *THUMOS: The First International Workshop on Action Recognition with a Large Number of Classes, in conjunction with ICCV '13, Sydney, Australia.*, 2013. 1, 2

[30] G. Willems, T. Tuytelaars, and L. V. Gool. An efcient dense and scale-invariant spatio-temporal interest point detector. In *ECCV*, 2008. 2

[31] L. Yeffet and L. Wolf. Local trinary patterns for human action recognition. In *ICCV*, 2009. 1, 2