Nested Motion Descriptors - Supplementary Material

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

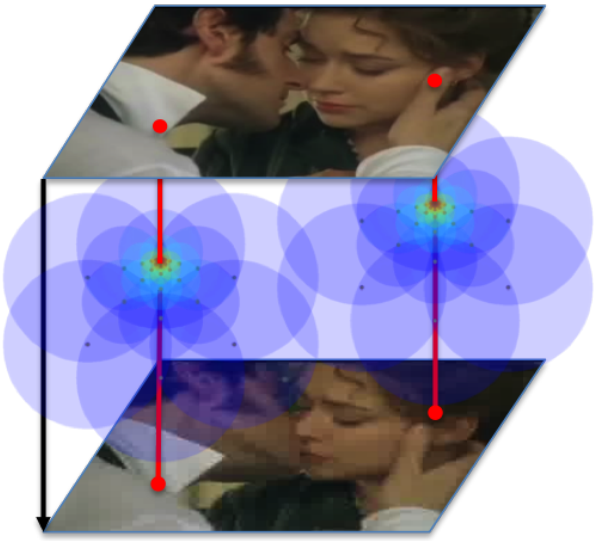


Figure 1. Nested motion descriptors - A Visualization

1. Complex Steerable Pyramid

The complex steerable pyramid [15, 14, 10] 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.

The complex steerable pyramid is computed using a recursive pyramid decomposition [10]. Given a set of steerable basis filters G and Hilbert transform H, let a basis filter F be represented in complex form by $F = G + H * i$. Each filter is tuned to a bandpass response in frequency ω and orientation θ forming a set of complex steerable filters $F_{\omega,\theta}$. The bandpass response $B_{\omega,\theta} = I \otimes F_{\omega,\theta}$ is formed by convolution of an image I with the complex filter. The pyramid decomposition is formed by recursively convolving an image I with a lowpass filter F_0, downsampling the image by a factor of 2, then computing the bandpass response B. This pyramid decomposition procedure is shown in figure 3 This decomposition can be made faster by considering separable kernels for the lowpass and complex steerable filters forming a separable quadrature steerable pyramid [14].

The complex steerable pyramid provides a measurement of the magnitude and phase of oriented and scaled edges. Following pyramid decomposition, complex valued bandpass coefficients can be decomposed into a real component representing the in-phase response, and the imaginary component representing the quadrature response. Let a coefficient $c_{\omega,\theta}(u,v) = x + iy$ be the complex valued coefficient for subband with orientation θ and scale ω for pixel (u,v) with real component x and imaginary component y. Then, the magnitude and phase of this coefficient is $|c| = \sqrt{x^2 + y^2}$ and phase of $\angle c = atan2(y,x)$. Intuitively, the magnitude is proportional to the contrast of an edge at the tuned orientation and scale at (u,v), and the phase is proportional to the shift in the direction of the tuned filter orientation to the dominant edge. In other words, phase encodes a *spatial offset* to an edge.

Figure 4 (a-d) shows an example of the magnitude and phase response of a complex quadrature filter for a 1D step edge. The impulse response of this real and imaginary component of this quadrature pair is shown in the second plot. Observe that these filters form a quadrature pair such that the quadrature component is shifted by $+\frac{\pi}{2}$ relative to the in-phase component. The phase plot show that the phase exhibits a *linear* response near the step edge (modulo π where the phase wraps from $+\pi$ to $-\pi$). Furthermore, the phase gradient is *constant* in this region and equal to one. This linearity of phase is exploited to estimate velocity in the next section.

2. Phase Gradients and Component Velocity

In general, the relationship between phase, translation and velocity is summarized in the *Fourier shift theorem*. This classic theorem states that a translation in the spatial

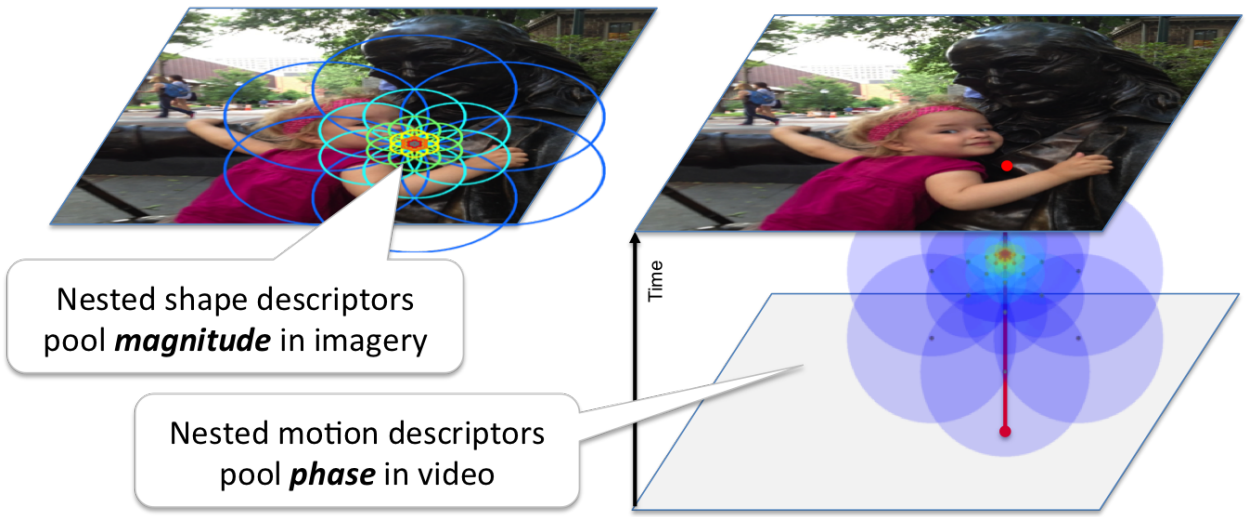


Figure 2. From nested shape descriptors to nested motion descriptors. Nested shape descriptors pool oriented and scaled gradients magnitude which captures the contrast of an edge in an image. Nested motion descriptors pool *relative phase* which captures *translation* of an edge. Projecting the structure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

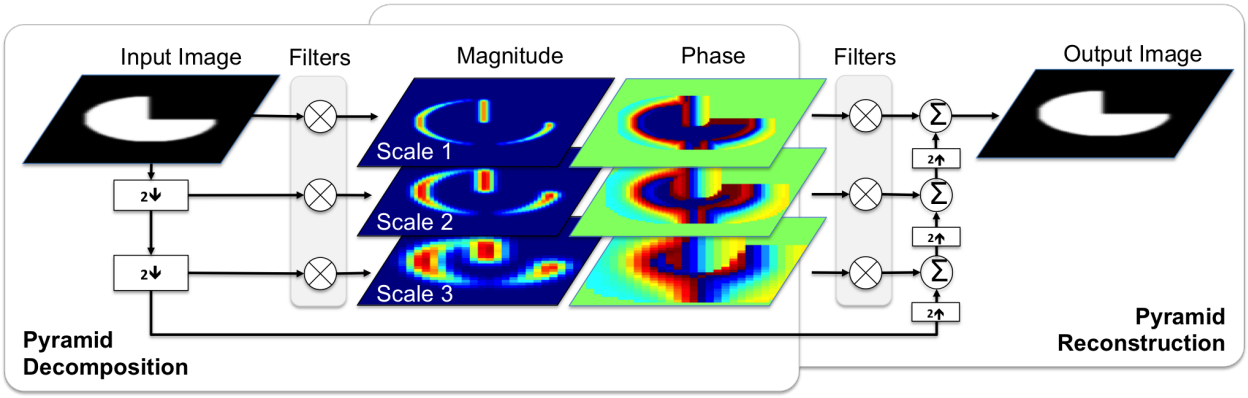


Figure 3. Pyramid decomposition and reconstruction with the complex steerable pyramid.

domain is equivalent to a phase shift in the frequency domain. In this section, we derive the relationship between phase and phase gradients to derive a measurement of velocity.

An interesting property of the complex steerable pyramid is the ability to introduce motion without changing position simply by varying the local phase. This phenomenon has been described as "Motion without Movement" [5], such that continuously varying the local phase of a bandpass response induces the visual phenomenology of global motion. This relationship between phase and motion has been used in phase based optical flow methods [3, 4] to enforce the *phase constancy* constraint [3], 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 [6, 4] over ranges of shape deformation and lighting. Recent work has exploited this relationship between phase and motion to amplify small changes in phase to visualize of microscopic motion at macro scale [17]. This approach multiplies small changes in phase by a large constant, then each image is reconstructed by collapsing the steerable pyramid, introducing a local image translation due to the local phase shift.

The phase constancy constraint is defined as follows [3]. Let a complex bandpass response B tuned to an orientation and scale be given by:

$$B(x,t) = \rho(x,t)e^{i\phi(x,t)} \tag{1}$$

The magnitude ρ and phase ϕ of this complex valued spatiotemporal function are also are spatiotemporal functions that evolve in space and time. Next, consider a moving point at x_0. This moving point evolves according to the *motion field*, a spatiotemporal vector field that defines the move-

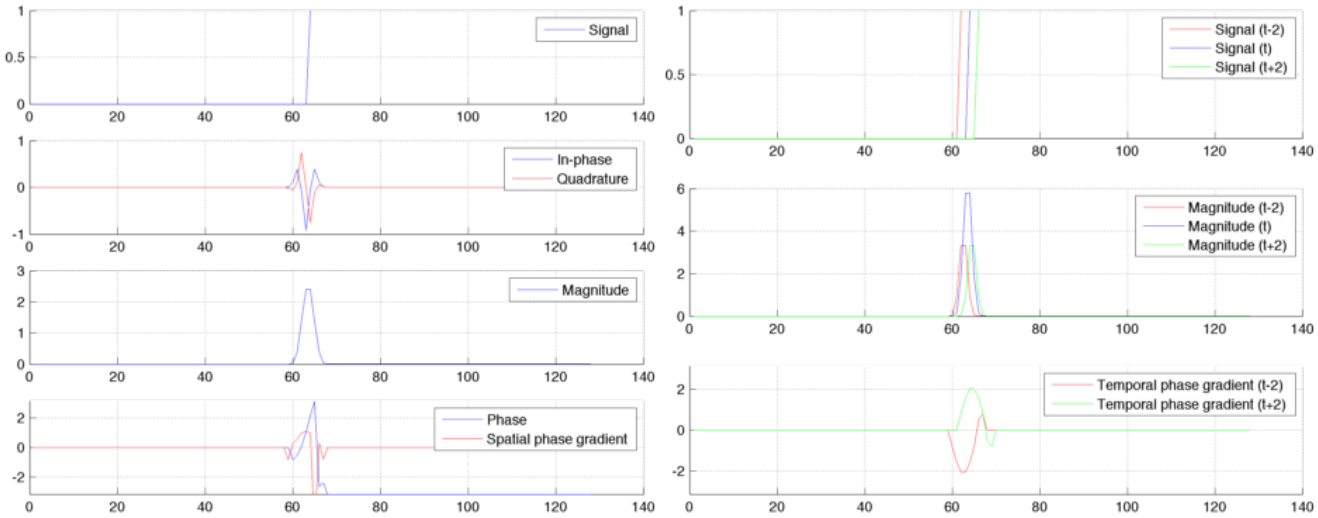


Figure 4. An example of the magnitude and phase response of a complex filter to a translating 1D step edge signal. (left column, top to bottom) (a) step edge signal (b) impulse response of 1D quadrature filters (c) magnitude of complex filter response to step edge (d) phase and spatial phase gradient ($|\vec{\phi}|$) of complex filter response showing linearity of phase. (right column, top to bottom) (e) A step edge translating left to right. (f) the magnitude response (g) the temporal phase gradient (ϕ_t). Observe that at the edge, the spatial phase gradient $|\vec{\phi}| = 1$ and the temporal phase gradient is $\phi_t = \pm 2$, which measures a spatial shift of $\frac{\phi_t}{|\vec{\phi}|} = \pm 2$.

ment of each pixel through time. The motion field is encoded as a function $x_0(t)$ which defines the spatial position of x_0 as a function of time. Fleet and Jepson in their seminal work on phase based optical flow [2, 6, 3, 4] hypothesized that the temporal evolution of spatial contours of constant phase provides a better approximation to the motion field than do contours of constant amplitude. This *phase contour assumption* states that the motion field must satisfy

$$\phi(x_0(t), t) = c \tag{2}$$

where c is a real valued constant. A point $x_0(t)$ propagating as a function of time according to the motion field is constrained to fall on a contour of constant phase $\phi(x_0(t), t)$. Intuitively, this states that phase is coherent and is preserved as a point propagates through time.

The phase contour assumption can be used to construct the *phase constancy constraint.* Differentiating the phase contour constraint, we obtain:

$$\bigtriangledown \phi(x, t) \bullet \vec{v} = 0 \tag{3}$$

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

$$\frac{\partial\phi}{dx}v_x + \frac{\partial\phi}{dy}v_y = -\frac{\partial\phi}{dt} \tag{4}$$

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. Observe that the dot product in (3) shows that the velocity cannot be determined normal to the phase gradient, which provides a constraint only on the component of velocity tuned to the orientation of the filter B. The phase constancy constraint in (4) 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 (4):

$$\alpha = \frac{-\phi_t}{|\vec{\phi}|} \tag{5}$$

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 $\vec{v} = \alpha\hat{n}$. Fleet and Jepson further proposed that the component velocities can be used as an overcomplete set of measurements to estimate the optical flow v using regularized least squares optimization. This can provide an estimate of pixel velocity or *optical flow* from measurements of

component velocity, which is the foundation of phase based optical flow methods [2, 6, 3, 4].

The component velocity (5) 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

$$\bigtriangledown \phi = \frac{Im(B^* \Delta B)}{|B|^2} \tag{6}$$

where $Im(z)$ is the imaginary component of the complex number z, and B^* is the complex conjugate of the complex valued bandpass response [3]. This identity for the phase gradient depends only on the complex bandpass response, and avoids an explicit computation of the phase angle using a trigonometric function.

Figure 4 (d-f) shows an example of the phase gradient and component velocity estimate. In this example, a 1D step edge is translating by two pixels from left to right. Figure 4 (e) shows the magnitude response of this translation, and (f) shows the temporal phase gradient computed using (6). The spatial phase gradient is shown in figure 4 (d). Using the measured phase gradients, we can use (5) to compute the velocity magnitude $\alpha = \frac{\pm 2}{1}$, which shows that the phase gradients provide a measurement of shift of the translating step edge.

3. Perspective views of the NMD

Figure 2 and figure 5 show perspective views of the nested motion descriptor. This provides a visualization of the 3D pooling structure of this descriptor.

4. Construction of the NMD

Figure 6 shows the log-spiral property the the log-spiral normalization of the nested motion descriptor.

5. Experimental Results

The goal of our experimental evaluation is demonstrating of *relative performance* of local motion descriptors for the task of activity recognition. This experimental evaluation does not attempt to achieve the state of the art in activity recognition on any one dataset. For example, the current state of the art uses higher level activity representations using improved dense trajectories and Fisher vector encoding of activities [19]. Instead, we are interested in determining the relative effect of only the local motion descriptors, in order to determine the relative benefit of this representation for this task. As a result, we consider only the relative performance of classification using a simple and well understood activity representation based on bag-of-words. This will not achieve state of the art, but the relative ranking is insightful for the performance of the descriptors only. These descriptors could then be used to improve the performance of dense trajectories to further push the state of the art. This evaluation strategy was used for baseline comparisons of local motion descriptors in activity recognition evaluations in [18, 1], and we follow the same approach.

We compare performance of the nested motion descriptors to HOG-HOF [9] and HOG-3D [7]. As described in the related work, there are many other motion descriptors including motion boundary histograms, motion interchange patterns and variants of dense trajectories. However, all of these descriptors are non-local. They focus on optical flow to aggregate local descriptors by tracking points through a long trajectory, which is a form of a global representation. In fact, dense trajectories define their representation as set of HOG-HOF descriptors extracted along a trajectory. The nested motion descriptor is local to a specific interest point, rather than capturing the properties of a trajectory. Therefore, we compare to other local motion descriptors. The evaluation in [18] 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. The KTH actions dataset [13] (2004), is representative of classic activity recognition dataset, with six classes and unmoving and zooming cameras. The UCF sports actions dataset [12] (2008) has nine activity classes, but these videos are collected in unconstrained television footage. Finally, the human motion database (HMDB) [8] (2011) is representative of a modern dataset with over fifty actions in unconstrained video.

The state of the art for activity recognition has moved to larger and more diverse datasets [11][16] with hundreds of activity classes, however since our focus is on relative performance of descriptors, we focus on classic datasets that span the complexity rather than pushing the absolute classification accuracy performance. Furthermore, classification performance has saturated on the KTH actions dataset to near perfect classification results, due primarily to the fact that the camera is not moving. However, remember that our analysis is focused on demonstrating the relative performance benefit of the local motion descriptors, and not the absolute classification performance of the activity recognition framework. So, these datasets remain informative for this relative analysis task.

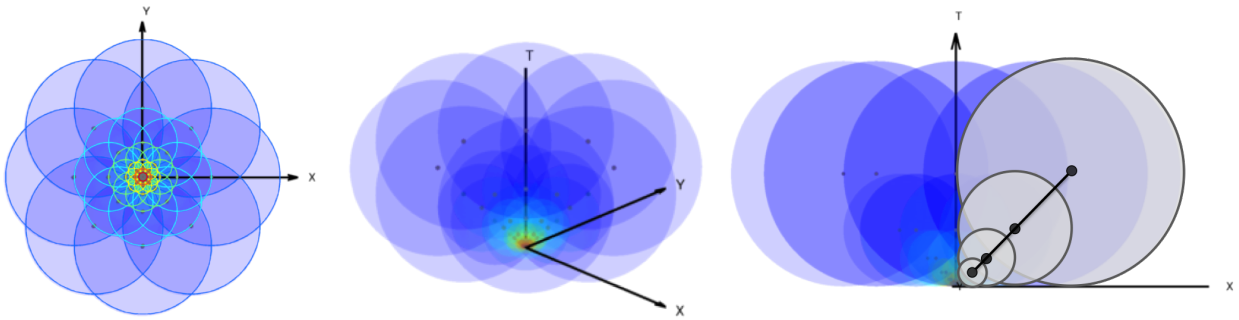


Figure 5. Perspective views of the spatiotemporal pooling regions of the nested motion descriptor. (left) az=90^o, el=90^o, the temporal axis is pointed into the page. We overlay the nested shape descriptor onto this view, which shows that the NMD has an equivalent pooling structure to the NSD (middle) az=45^o, el=25^o, with the temporal axis pointed into the page, (right) az=90^o, el=0^o, with the Y axis pointed out of the page. This view shows that the temporal pooling regions increase proportionally to spatial scale. The slope of the line connecting the centers is determined by the velocity tuning of the descriptor. A video visualization of this descriptor is available in the supplementary material.

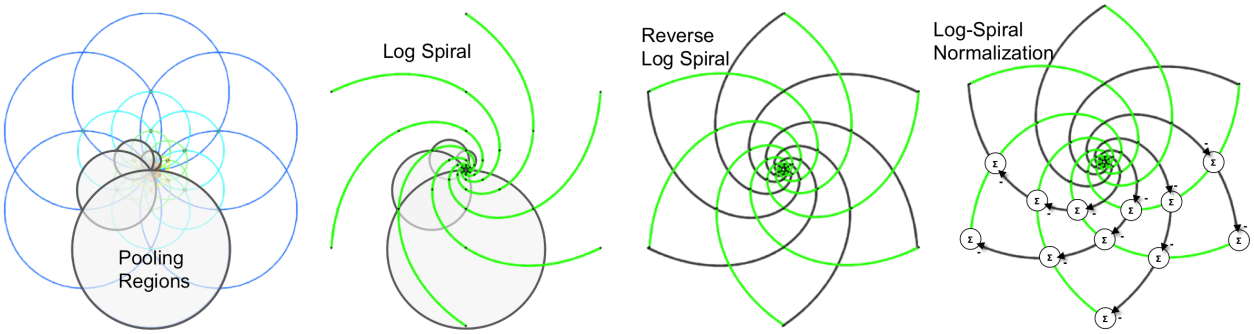


Figure 6. (top) Logarithmic spiral property of the nested motion descriptor provides *normalization* and *binarization*. The log-spiral and it's reflection shown in grey form an elegant flower-like structure. (bottom) An NMD is formed at each interest point by (left) nested pooling of scaled and oriented gradients and (right) log-spiral difference and binarization.

5.1. Experimental System

The experimental system we consider for evaluation of nested motion descriptor performance is activity recognition using a bag of words representation.

For each observation, we densely extract local motion descriptors from each frame in the video, with the given spatiotemporal stride. We use a all descriptors from a random sample of 30 videos to perform vector quantization to learn the K words in the vocabulary. Then, for each video, we construct a bag-of-words representation by assigning each densely extracted descriptor to the closest word, and creating a normalized histogram of word occurrence. Finally, classification is performed by training a one-vs-rest linear SVM classifier for each class, then selecting the maximum likelihood class for each observation. We report results in classification rate or mean average precision across all classes for each dataset.

We compare to the baseline of [9] and HOG-3D [7] local motion descriptors. We use the public implementations available from the author's websites, and initialize these descriptors to the parameters listed below.

Finally, we use the following parameters in all experiments, in addition to the default parameters recommended by the original authors.

- **Resolution:** We downsample frames so that the maximum dimension is 160 pixels.
- **Visual words:** 600 words in the vocabulary, trained from a random sample of 10,000 descriptors from 30 videos.
- **Stride:** dx=5, dy=5 spatially, dt=5 temporally
- **NMD parameters:** scales=5, orientations=8, lobes=8, real valued (without binarization), with log-spiral normalization
- **Dataset size per class:** 30 training videos, 65 testing videos.

Training and testing splits follow the recommendations from the dataset authors, unless otherwise noted. For KTH actions, we follow the recommended training and testing splits where we divide the test set into nine subjects (2, 3, 5, 6, 7, 8, 9, 10, and 22) and the training set into the remaining subjects. For HMDB, we use the unstabilized HMDB

videos and limit the training and testing to the listed number of videos per class above. For UCF sports, we perform leave one out cross validation due to the limited number of videos available per class and report only confusion matrix and mean classification rate results.

5.2. Experimental Datasets

KTH actions is a classic activity recognition dataset [13]. This dataset contains six types of human actions (walking, jogging, running, boxing, hand waving and hand clapping) performed several times by 25 subjects in four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes and indoors. This dataset contains 2391 sequences, such that all sequences were taken over homogeneous backgrounds with a static camera with 25 Hz frame rate. The sequences were downsampled to the spatial resolution of 160x120 pixels and have an average length of four seconds.

The **UCF sports actions dataset** [12] consists of a set of nine actions collected from various sports typically featured on broadcast television channels such as the BBC and ESPN. The video sequences were obtained from a wide range of stock footage websites including BBC Motion gallery, and Getty Images. This dataset contains close to 200 video sequences at a resolution of 720x480. The collection represents a natural pool of actions featured in a wide range of scenes and viewpoints. Actions in this data set include: Diving (16 videos), Golf swinging (25 videos), Kicking (25 videos), Lifting (15 videos), Horseback riding (14 videos), Running (15 videos), Skating (15 videos), Swinging (35 videos) and Walking (22 videos).

The **Human Motion DataBase (HMDB)** is a recent activity dataset containing a large number of activities in the wild [8]. HMDB is an activity recognition dataset collected from various sources, mostly from movies, and a small proportion from public databases such as the Prelinger archive, YouTube and Google videos. The dataset contains 6849 clips divided into 51 action categories, each containing a minimum of 101 clips. The categories can be grouped in five types:

- General facial actions such as smile, laugh, chew, talk.
- Facial actions with object manipulation such as smoke, eat, drink.
- General body movements: cartwheel, clap hands, climb
- Body movements with object interaction: brush hair, catch, draw sword,
- Body movements for human interaction: fencing, hug, kiss

6. Additional Motion Visualization Examples

Figure 7 shows an example of jogging from the KTH actions dataset. This example considers a static camera, so there is zero motion in the background due to camera motion. The bottom row shows the motion visualization without the log-spiral normalization, and this shows that the motion is dominated by the overall movement of the jogger from right to left. The top row shows the effect of the log-spiral normalization which causes the motion of the legs and pumping of the arms to pop out.

Figure 8 shows an example of hug from the HMDB. This example also includes a camera motion panning from left to right as the two people converge to a hug. Without the log-spiral normalization, this camera motion dominates, reducing the scene to a single motion blob. With the log-spiral normalization, the salient motion of the hands and head as two enter the hug.

7. KTH Actions

Figure 9 shows detailed classification results on the KTH actions dataset. This dataset has a large number of training examples per class which allows for evaluation using precision-recall curves in addition to the confusion matrices and classification rates. This result shows that the NMD exhibits significantly improved average precision for boxing and handwaving, but is worse on jogging. An analysis of the confusion matrix for the NMD shows that performance on jogging is confused with running and walking. This suggests that the absolute velocity is a discriminative feature for this class, and the log-spiral normalization discards this information when computing the invariance to camera motion. It is interesting to note that in some cases, the dominant motion in the scene is informative for classification. This highlights the need for a composition of various descriptors for accurate activity classification.

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. 4

[2] D. Fleet. *Measurement of Image Velocity*. Kluwer Academic Press, 1992. 3, 4

[3] 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

[4] 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

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

[6] A. Jepson and D. Fleet. Phase singularities in scale-space. *Image and Vision Computing Journal*, 9(5):338–343, 1991. 2, 3, 4

[7] A. Klaser, M. Marsza?ek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, 2008. 4, 5

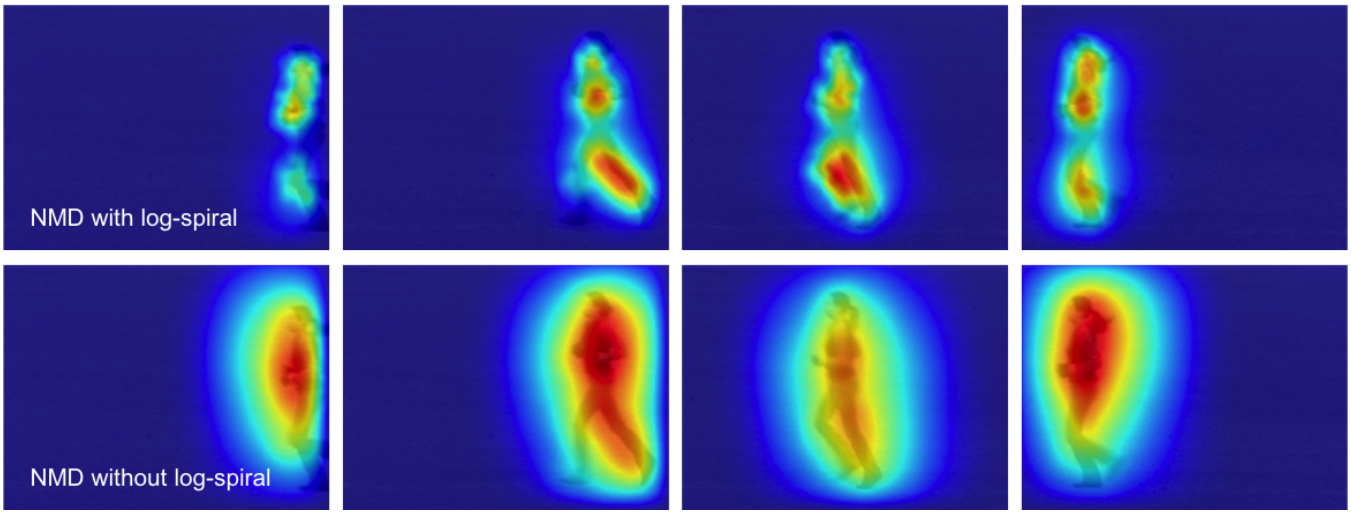


Figure 7. Motion saliency for KTH jogging. (top) Salient motion using NMD with log spiral normalization. (bottom row) NMD without log-spiral normalization. The log-spiral normalization highlights the salient motion of the runners legs and arms, while the motion without log-spiral normalization saturates with the motion of the mean velocity of the body. A video visualization of this motion saliency is provided in the supplementary material.

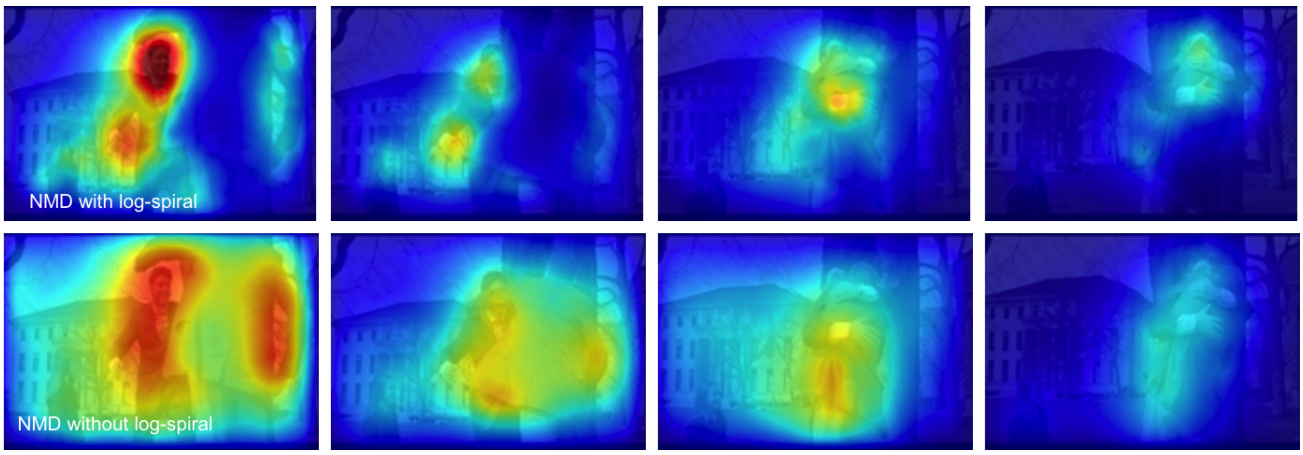


Figure 8. Motion saliency for HMDB hug. (top) Salient motion using NMD with log spiral normalization. (bottom row) NMD without log-spiral normalization. The log-spiral normalization focuses on the subtle hand movements that form a hug and suppresses the background motion of the camera. A video visualization of this motion saliency is available in the suplementary material.

[8] H. Kuhne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition. In *ICCV*, 2011. 4, 6

[9] I. Laptev. On space-time interest points. *IJCV*, 2005. 4, 5

[10] J. Portilla and E. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, 2000. 1

[11] K. K. Reddy and M. Shah. Recognizing 50 human action categories of web videos. *Machine Vision and Applications Journal (MVAP)*, 2012. 4

[12] M. Rodriguez, J. Ahmed, and M. Shah. Action mach: A spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, 2008. 4, 6

[13] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *ICPR*, 2004. 4, 6

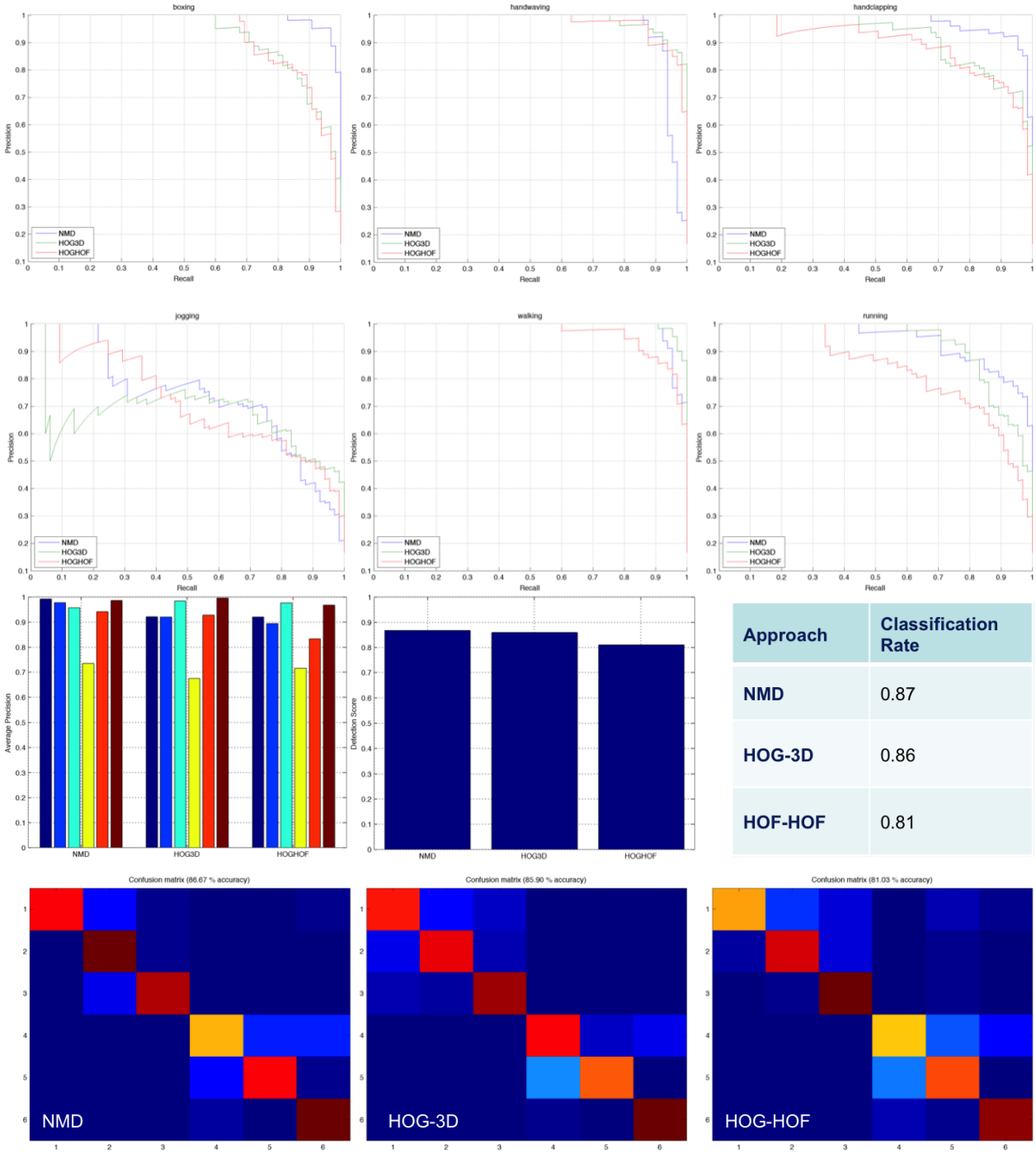
[14] E. Simoncelli and W. Freeman. The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *IEEE Second Int'l Conf on Image Processing*, 1995. 1

[15] E. Simoncelli, W. Freeman, E. Adelson, and D. Heeger. Shiftable multi-scale transforms. *IEEE Trans. Info. Theory*, 2(38):587–607, 1992. 1

[16] K. Soomro, A. Roshan, and M. Shah. Ucf101: A dataset of 101 human action classes from videos in the wild. Technical Report CRCV-TR-12-01, UCF, November 2012. 4

[17] 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

[18] H. Wang, M. M. Ullah, A. Klser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009. 4



Approach	Classification Rate
NMD	0.87
HOG-3D	0.86
HOF-HOF	0.81

Figure 9. Activity classification results on KTH actions (top) Precision-recall curves for each of six activity classes (middle) average precision per class, mean classification rate, (bottom) confusion matrices. For all results, the class indexes are ordered: boxing, handclapping, handwaving, jogging, running, walking. NMD results are improved for boxing and handclapping, but worse for jogging. See text body for a discussion.

[19] 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. 4