Nested Motion Descriptors

Jeffrey Byrne[1,2]
[1]University of Pennsylvania, GRASP Lab [2]Systems and Technology Research

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 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". 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 [6].

Recent work has focused on motion descriptors that are invariant to camera motion [5, 6, 8, 10, 11, 12, 13, 14]. Local spatiotemporal descriptors such as, such as HOG-HOF [2, 9] or HOG-3D [7], 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, which can corrupts the motion representation.

In this paper, we propose a new family of binary local motion descriptors called *nested motion descriptors*. 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 [1] to the representation of motion. 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 [4], component velocity [3] and motion without movement, which uses phase shifts as a correction for translation without an explicit motion field estimate.

This paper demonstrates 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.

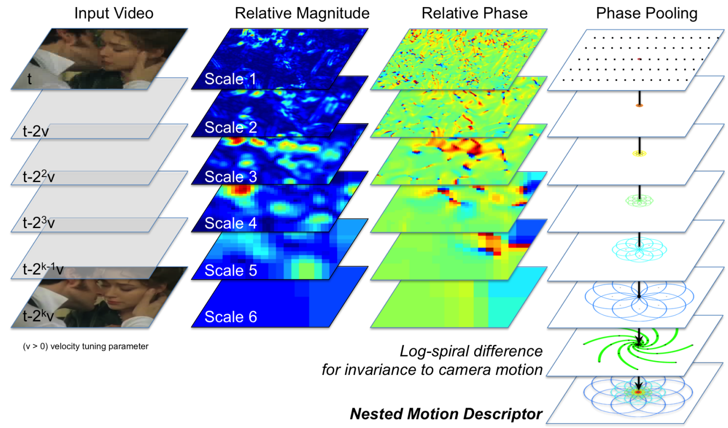

Figure 1: Nested Motion Descriptors (NMD). (left) Compute relative magnitude and phase for orientations and scales for a set of frames, (right) Pool the robust component velocity derived from relative phase in a set of circular pooling regions all intersecting at the center interest point. Log-spiral 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] J. Byrne and J. Shi. Nested shape descriptors. In *ICCV*, 2013.

[2] Navneet Dalal, Bill Triggs, and Cordelia Schmid. Human detection using oriented histograms of flow and appearance. In *ECCV*, 2006.

[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.

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

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

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

[7] Alexander Klaser, Marcin Marsza?ek, and Cordelia Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, 2008.

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

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

[10] 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.

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

[12] H. Wang, A. Klï£¡ser, C. Schmid, and L. Cheng-Lin. Action recognition by dense trajectories. In *CVPR*, 2011.

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

[14] 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.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.