Nested Motion Descriptors

Xuewen Yang

June 16 2018

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 [1] 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 HOGHOF, and show improvements on three activity recognition datasets

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 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 [4]. 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 [2]. Recent work has focused on motion descriptors that are invariant to camera motion [2]. Local spatiotemporal descriptors such as, such as HOG-HOF or HOG-3D [3], 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. 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, component velocity and motion without movement, 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.

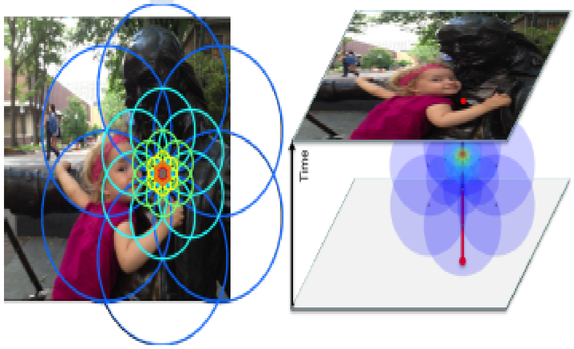

Figure 1. From nested shape descriptors to nested motion descriptors. Nested shape descriptors pool oriented and scaled gradients magnitude which captures the contrast of an edge in an image. Nested motion descriptors pool relative phase which captures translation of an edge. Projecting the spatiotemporal structure of the nested motion descriptor onto a single image will form the structure of the nested shape descriptor.

Examples of local motion descriptors include HOG-HOF, cuboid, extended SURF and HOG-3D [3]. 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 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, 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 [2], and removing the effect of this global motion significantly improves the representation of foreground motion for activity recognition. The motion boundary histogram 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 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 concatenate HOG-HOF and motion boundary histograms for a tracked sequence of interest points forming a long term trajectory descriptor. The improved dense trajectories with fisher vector encoding is the current state-of-the-art on large datasets for action recognition.

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 [1] to the representation of motion. Figure1 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.

References

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

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

[3] A. Klaser. A spatiotemporal descriptor based on 3D-gradients. In *BMVC*, 2008. 1, 2

[4] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, 2009. 1