Finding Actions Using Shape Flows

Hao Jiang and David R. Martin

Computer Science Department, Boston College, Chestnut Hill, MA 02467, USA
{hjiang,dmartin}@cs.bc.edu

Abstract. We propose a novel method for action detection based on a new action descriptor called a *shape flow* that represents both the shape and movement of an object in a holistic and parsimonious manner. We find actions by finding shape flows in a target video that are similar to a template shape flow. Shape flows are largely independent of appearance, and the match cost function that we propose is invariant to scale changes and smooth nonlinear deformation in space and time. The problem of matching shape flows is difficult, however, yielding a large, non-convex, integer program. We propose a novel relaxation method based on *successive convexification* that converts this hard program into a vastly smaller linear program: By using only those variables that appear on the 4D lower convex hull of the matching cost volume, most of the variables in the linear program may be eliminated. Experiments confirm that the proposed shape flow method can successfully detect complex actions in cluttered video, even with self-occlusion, camera motion, and intra-class variation.

1 Introduction

An action can be characterized by the movement and deformation of a shape. A *flow line*, which is the space-time line formed by a tracked object point over time, provides a compact representation of motion. An object's *shape flow*, which we define simply as an assembly of flow lines, represents not only the object's motion but also its shape and deformation over time. Flow lines have previously been used for motion visualization [5]. We propose the shape flow as a representation for actions.

Consider the shape flows shown in Fig. 1. One can readily identify complex actions based on the shape flow alone, suggesting that it is a discriminative and descriptive representation. The shape flow itself is quite simple to compute. The challenge, which is the focus of this paper, is how to use the shape flow as a representation for actions. In particular how can we efficiently search for a template shape flow in a cluttered single-view video? And how can we do the search in a manner that is invariant to scale changes and nonlinear shape deformations, and also tolerant of occlusion and intra-class variation? Although shape flow matching is certainly an NP-hard problem because of the loopy relations between flow lines, we show that a novel linear relaxation based on *successive convexification* [6] yields both an efficient and accurate matching algorithm for finding actions in video.

There is much related work on detecting actions in video. The first dimension of related work consists of methods that use multi-view stereo to reliably access 3D spatial information. Parameswaran & Chellappa [12] use multiple cameras to capture joint

D. Forsyth, P. Torr, and A. Zisserman (Eds.): ECCV 2008, Part II, LNCS 5303, pp. 278–292, 2008.

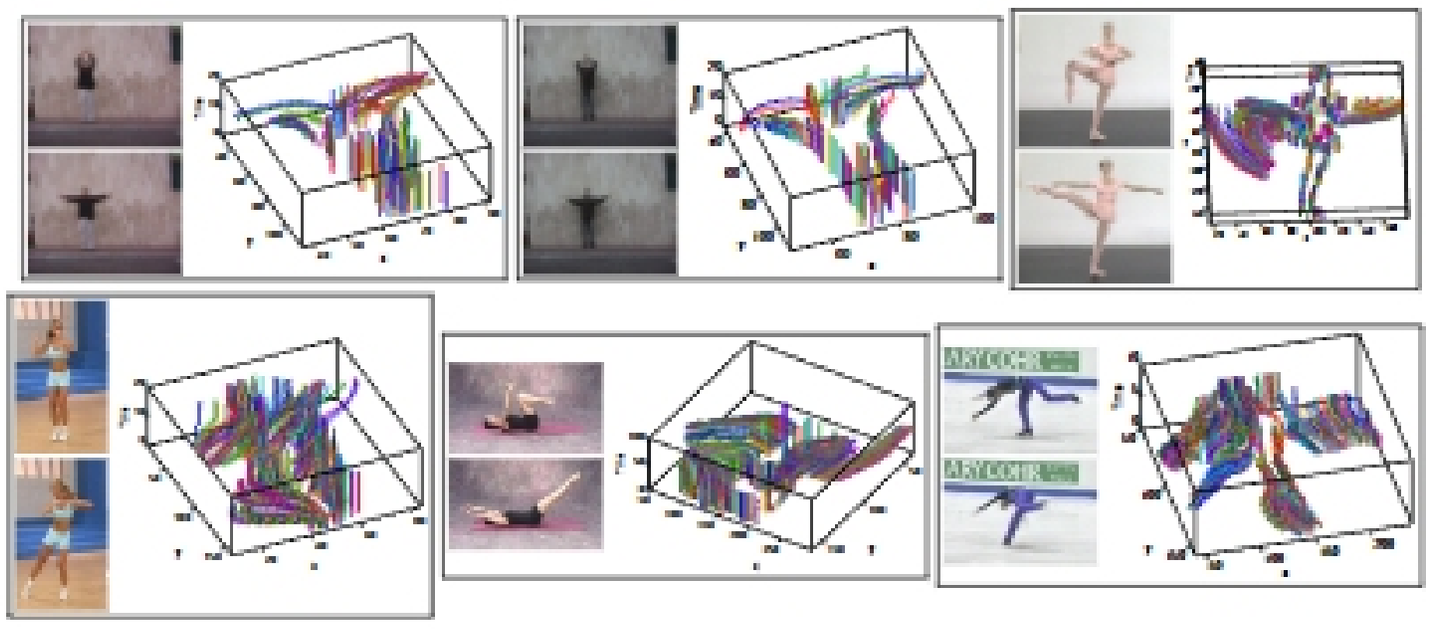

Fig. 1. Example shape flows and first/last frames for a variety of actions. Even though individual flow lines are noisy, the shape flow represents the holistic shape *and* movement of the object reliably. In this paper, we show how to efficiently search for a template shape flow in a target video. *Please note that the figures in this paper are best viewed in color.*

locations via motion capture, while Yilmaz & Shah [20] use manually labeled joint locations. Weinland et al. [18] use multiple cameras to achieve reliable background subtraction, using 3D silhouettes of foreground objects to describe actions. These methods are only tangentially related to this paper, however, as they rely on multi-view stereo.

The second dimension of related work includes methods that rely on background subtraction to formulate action features from silhouettes of foreground objects either in 2D or in 3D. Weinland et al. [18] use 3D silhouettes from multi-view stereo, as mentioned above. Both Yilmax & Shah [19] and Blank et al. [2] use background subtraction in 2D images to describe actions as 3D space-time volumes. Bobick & Davis [3] project space-time silhouettes into the image to get 2D action silhouettes. In this paper, we address the problem of finding actions in cluttered single-view video with a moving camera and moving background objects. Silhouette based features, which rely on high quality background subtraction, are difficult to use in this regime.

The third dimension of related work involves using labeled joint trajectories to detect human actions. In this class-specific approach, joint locations may be labeled in different ways. Parameswaran & Chellappa use motion capture and manual labeling for joint locations in 2D and 3D in their work [12]. Yilmaz & Shah [20] use manually labeled joint trajectories in 3D, and Sheikh & Shah [17] use manually labeled joint trajectories in 2D. Recognizing actions is challenging even with clean joint trajectories. Although automatic human pose tracking is recently much improved [11,13] and so could be used to extract (unclean) joint trajectories from unlabeled video, we pursue a non-parametric approach of analyzing the whole motion field rather than the motion of distinguished high-level feature points.

The final dimension of related work consists of non-parametric action models, and contrasts methods that rely on sparse descriptors located at interest points versus methods that use a dense motion or gradient field. These methods typically operate on uncalibrated single-view video, do not require background subtraction, and do not require

manual labeling of foreground objects. Laptev & Lindeberg [9] have extended the notion of Harris extrema to find space-time interest points. Schuldt et al. [14] use motion and gradient histograms around these interest points to recognize actions using SVMs. Scovanner et al. [15] extend the popular SIFT descriptor to space-time volumes for the purpose of action recognition. Instead of constructing features that are tolerant to clutter, one may instead use dense feature fields along with clutter-tolerant matching. Efros et al. [4] recognize actions using optical flow fields from single frames, carefully smoothed and rectified. Shechtman & Irani [16] cleverly match space-time volumes directly, without any explicit motion estimate. Ke et al. [7] have extended that work using superpixels and part-based matching. Laptev & Prez [10] match histograms of gradients from space-time cubes using a boosted cascade.

These non-parametric methods represent a recent trend toward using volumetric space-time descriptors that combine shape *and* motion information, along with a matching framework that tolerates both foreground deformation and background clutter. We adopt this general approach to finding actions, although our proposed shape flow descriptor is neither an interest point method nor a dense field approach. As is clear in Fig. 1, the shape flow preserves much information about both the shape and movement of an object in a manner that is largely independent of the object's appearance. This enables us to match both shape and motion in a uniform framework.

The outline of the paper is as follows. In §2 we describe a simple method to compute shape flows. In §3, we formulate the shape flow matching problem and show how this NP-hard problem may be solved efficiently using a novel relaxation scheme. We present experimental results in §4 and conclude in §5.

2 The Shape Flow of an Action

We desire a compact yet expressive descriptor of both the shape and motion of an object. In addition, we seek a general method that may be applied to any object class, so we do not attempt to track distinguished points (such as joints) or to impose a shape model a priori. Instead, we seek a non-parametric shape and motion representation. We adapt the technique of flow fields from motion visualization [5]. If points on an object may be tracked in 2D video, then their positions through time form a flow field of lines in 3D. Flow lines will inevitably be individually unreliable, but the collection of flow lines—the *shape flow*, as shown in Fig. 1—is a compact and descriptive action representation.

We use a greedy scheme to compute flow lines based on iterative conditional modes (ICM) [1] to estimate the sparse point motion between adjacent frames. ICM is based on an MPEG-like local motion estimation search; it computes motion vectors that both minimize an image match cost and maximize the the motion consistency of neighboring points. We apply ICM to edge pixels that surpass a Canny detector threshold; neighborhood relations are defined by the Delaunay triangulation of these edge pixels. The resultant sparse motion field is then interpolated across Delaunay cells to produce a dense motion field. The frame-by-frame motion fields produced by this procedure are then simply concatenated to form 3D flow lines in the space-time video volume. There are no constraints to prevent flow lines from intersecting.

This method for computing flow lines is designed to produce flow lines that are good enough on average to generate a coherent flow field, since our robust matching