Articulated Motion Discovery using Pairs of Trajectories

Luca Del Pero,[1] Susanna Ricco,[2] Rahul Sukthankar,[2] Vittorio Ferrari[1]
[1]University of Edinburgh [2]Google Research

We propose a bottom-up, unsupervised method for discovering the characteristic motion patterns of a highly-articulated object class *in the wild*. Unlike the majority of action recognition datasets, in which human actors perform scripted actions, and/or clips are trimmed to contain a single action, our videos are unstructured, such as animals performing unscripted behaviors. The only assumption we make is that each video contain at least one instance of the object class. We leverage that the object is engaged in some (unknown) behaviors, and that such behaviors exhibit observable consistency, which we term characteristic *motion patterns*.

Our method does not require knowledge of the number or types of behaviors, nor that instances of different behaviors be temporally segmented within a video. The output of our method is a set of video intervals, clustered according to the observed characteristic motion patterns. Each interval contains one temporally segmented instance of the pattern. Fig. 1 shows some behaviors automatically discovered in tiger videos, such as walking, turning head, and running.

We identify consistency between observed motion patterns by analyzing the relative displacement of large numbers of ordered trajectory pairs (PoTs). The first trajectory in the pair defines a reference frame in which the motion of the second trajectory is measured. We preferentially sample trajectory pairs across joints, resulting in features particularly well-suited to representing fine-grained behaviors of complex, articulated objects. This has greater discriminative power than state-of-the-art features defined using single trajectories in isolation.

In contrast to other popular descriptors (e.g., [2]), PoTs are appearance-free. They are defined solely by motion and so are robust to appearance variations within the object class. In cases where appearance proves beneficial for discriminating between behaviors of interest, it is easy to combine PoTs with standard appearance features.

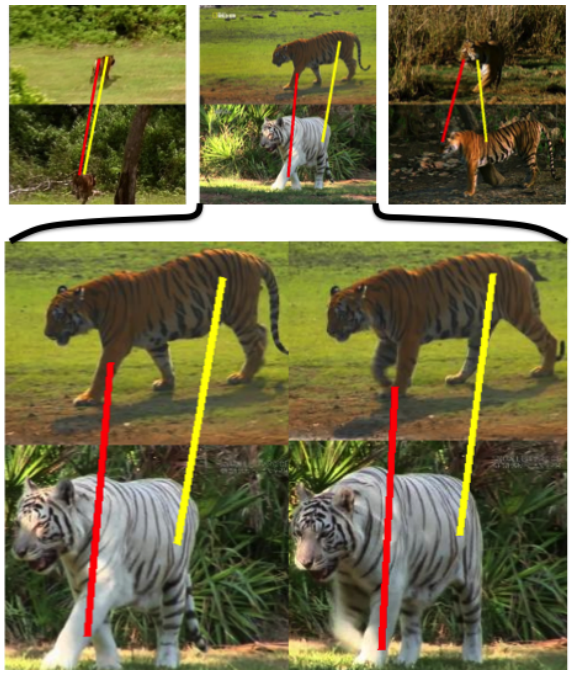

Figure 1: Examples of articulated motion pattern clusters discovered using pairs of trajectories (PoTs). These clusters capture tigers running, walking, and turning their heads, respectively. Inset shows detail of one PoT within the walking cluster. Yellow lines connect the first trajectories (on the tigers' body); red lines connect the second (on moving extremities).

Anatomy of a PoT

Anchors and swings. The first trajectory in each PoT (termed the *anchor*) defines a local coordinate frame, in which the motion of the second (termed the *swing*) is measured. Intuitively, the anchor is the trajectory in the pair that moves closer to the median velocity of the object while the swing is free to articulate. For example, in Fig. 1, trajectories on the legs would be chosen as swings while those on the torso would be anchors.

Displacement vectors. We capture the motion of the swing relative to the anchor by computing the *change* in the displacement vector from anchor to swing in consecutive frames (denoted as $\mathbf{d^k}$), over a temporal interval.

PoT descriptor. The PoT descriptor P consists of two parts: 1) the initial position of the swing relative to the anchor and 2) the sequence of normalized displacement vectors through time:

$$P = \left(\theta, \frac{\mathbf{d^2}}{D}, \ldots, \frac{\mathbf{d^n}}{D}\right), \quad (1)$$

where θ is the angle from anchor to swing in the first frame and the normalization factor is the total displacement $D = \sum_{k=2}^{n} ||\mathbf{d^k}||$. The DTF descriptor [2] employs a similar normalization. Note also that the first term in P records only the angle (and not the magnitude) between anchor and swing; this retains scale invariance and enables matching PoTs from objects of different size.

Evaluation

Datasets. We experiment on two different datasets. First, we use a dataset of tiger videos collected from National Geographic documentaries. This dataset contains roughly two hours of high-resolution, professional footage divided into 500 shots, for a total of 80,000 frames. Second, we use 100 shots of the dog class of the YouTube-Objects dataset [1], which mostly contains low-resolution footage filmed by amateurs.

Behavior labels. We annotated *each frame* in the dataset independently, from a set of animal behaviors (23 tiger behaviors and 15 dog behaviors). When a frame shows multiple behaviors, we chose the one that happens at the larger scale (e.g., we choose "walk" over "turn head" and "turn head" over "blink").

Results. We use two criteria commonly used for evaluating clustering methods: *purity* and *Adjusted Rand Index* (ARI). The clusters found using PoTs are better in both purity and ARI than those found using the state-of-the-art Improved Dense Trajectory Features (IDTFs) [2] on the tiger dataset, even though PoTs incorporate no appearance information. On the dog dataset, IDTFs perform slightly better than PoTs until we augment our representation with appearance information (PoTs+HOG).

Summary

Our contributions are: (1) a new feature based on ordered pairs of trajectories that captures the intricate motion of articulated objects; (2) a method for unsupervised discovery of behaviors from unconstrained videos of an object class; (3) a method to segment video into temporal intervals likely to contain single behaviors; and (4) annotations for 80,000 frames from nature videos about tigers and 20,000 from YouTube videos of dogs, available on our website: http://groups.inf.ed.ac.uk/calvin/proj-pots/page/

[1] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari. Learning object class detectors from weakly annotated video. In *CVPR*, 2012.

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

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