Dynamically Encoded Actions based on Spacetime Saliency

Christoph Feichtenhofer[1], Axel Pinz[1], Richard P. Wildes[2]
[1]Inst. of Electrical Measurement and Measurement Signal Processing, TU Graz, Austria.
[2]Department of Electrical Engineering and Computer Science, York University, Toronto, Canada.

Human actions typically occur over a well localized extent in both space and time. Similarly, as typically captured in video, human actions have small spatiotemporal support in image space. This paper capitalizes on these observations by weighting feature descriptors for action recognition over those areas within a video where actions are most likely to occur. To enable this selective pooling operation, we define a novel measure of spacetime saliency, $\mathcal{S}_\mathrm{T}$, that is designed to highlight human actions in temporal sequences of images. Figure 1 illustrates our approach for a ball catching action and compares our spacetime saliency measure with ground truth annotations from [3] and [9].

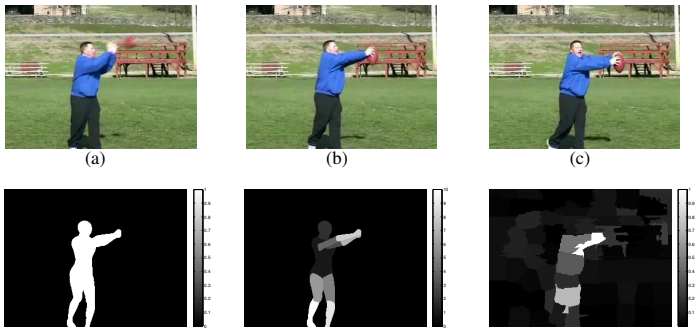
In our work, we adopt the popular Bags-of-visual-Word (BoW) approach to recognition consisting of three general steps: primitive feature extraction, feature encoding and feature pooling that accumulates encoded features over pre-defined regions. The major novelty in our contribution arises via selective encoding and pooling. Our approach dynamically encodes and pools primitive feature measurements via a new definition of spacetime saliency weights, $\mathcal{S}_\mathrm{T}$. While previous research has made use of saliency measures as part of an action recognition approach (e.g., [1, 7]), ours is novel in its definition of spacetime saliency based on directional motion energy contrast and spatial variance to capture actions. These two components are motivated directly by two corresponding observations about foreground human actions. First, actions typically involve a foreground motion that is distinct from the surrounding background. Indeed, even in the presence of global camera motion, a foreground action will exhibit a different (superimposed) pattern of motion. For example, a participant in a sporting event will yield a motion that is distinct from that of overall camera motion when he or she is engaged in their sporting activity. We refer to this property as *motion contrast*. Second, action patterns typically are spatially compact, while background motions are more widely distributed across an image sequence. For example, even interactions between two people (e.g., a hug, handshake or kiss) occupy a relatively small portion of an image. We refer to this second property as *motion variance*. In combination, these two properties are used to define our measure of spacetime saliency, $\mathcal{S}_\mathrm{T}$, for capturing foreground action motion.

In general, our dynamically adaptive saliency weighting can be applied essentially to any local feature measurements. Here, it is illustrated using various combinations of improved dense trajectories (IDT) [8] and a novel extension to Spatiotemporal Oriented Energies (SOE) [2], termed Locally Aggregated Temporal Energy (LATE), both encoded by Fisher vectors [5]. Significantly, in empirical evaluation the approach is competitive with and can even exceed the previous state-of-the-art in action recognition on three standard datasets, including J-HMDB [3], HMDB51 [4] and UCF101 [6]. Table 1 provides a sampling of our empirical results, with detailed comparison to alternatives available in the complete paper.

The results of our paper suggest the importance of explicitly concentrating processing on regions where an action is likely to occur during recognition.

Dataset	Accuracy
J-HMDB [3]	65.9
HMDB51 [4]	62.2
UCF101 [6]	87.7

Table 1: Mean classification accuracy of proposed approach over three train/test splits on J-HMDB, HMDB51 and UCF101. Detailed comparison to alternative approaches is provided in the complete paper.



(a) (b) (c) (d) GT-mask (e) GT-flow (f) $\mathcal{S}_\mathrm{T}$

Figure 1: Our proposed measure of spacetime saliency, $\mathcal{S}_\mathrm{T}$, to enhance feature pooling for action recognition. Input frames of a video from the J-HMDB dataset [3] showing a catch action. (d) Ground truth puppet mask annotation [3]. (e) The puppet flow generated from ground truth body part annotations [9]. (f) Our spacetime saliency measure, $\mathcal{S}_\mathrm{T}$, for weighting the contribution of local spatiotemporal features for action recognition. Note the large similarities between the user-annotated puppet flow and our saliency measure which is computed efficiently from motion statistics without any form of supervision.

Acknowledgments

This work was supported by the Austrian Science Fund (FWF) under project P27076 and a Canadian NSERC Discovery Grant.

References

[1] Nicolas Ballas, Yi Yang, Zhen-Zhong Lan, Bertrand Delezoide, Francoise Preteux, and Alexander Hauptmann. Space-time robust representation for action recognition. In *Proc. ICCV*, 2013.

[2] Konstantinos G. Derpanis, Mikhail Sizintsev, Kevin J. Cannons, and Richard P. Wildes. Action spotting and recognition based on a spatiotemporal orientation analysis. *PAMI*, 35(3):527–540, 2013.

[3] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J. Black. Towards understanding action recognition. In *Proc. ICCV*, 2013.

[4] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. HMDB: a large video database for human motion recognition. In *Proc. ICCV*, 2011.

[5] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the Fisher kernel for large-scale image classification. In *Proc. ECCV*, 2010.

[6] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions calsses from videos in the wild. Technical Report CRCV-TR-12-01, UCF Center for Research in Computer Vision, 2012.

[7] Waqas Sultani and Imran Saleemi. Human action recognition across datasets by foreground-weighted histogram decomposition. In *Proc. CVPR*, 2014.

[8] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proc. ICCV*, 2013.

[9] Silvia Zuffi and Michael J Black. Puppet flow. Technical Report TRIS-MPI-007, MPI for Intelligent Systems, 2013.

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