# Pooled Motion Features for First-Person Videos

M. S. Ryoo, Brandon Rothrock, Larry Matthies
Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA.

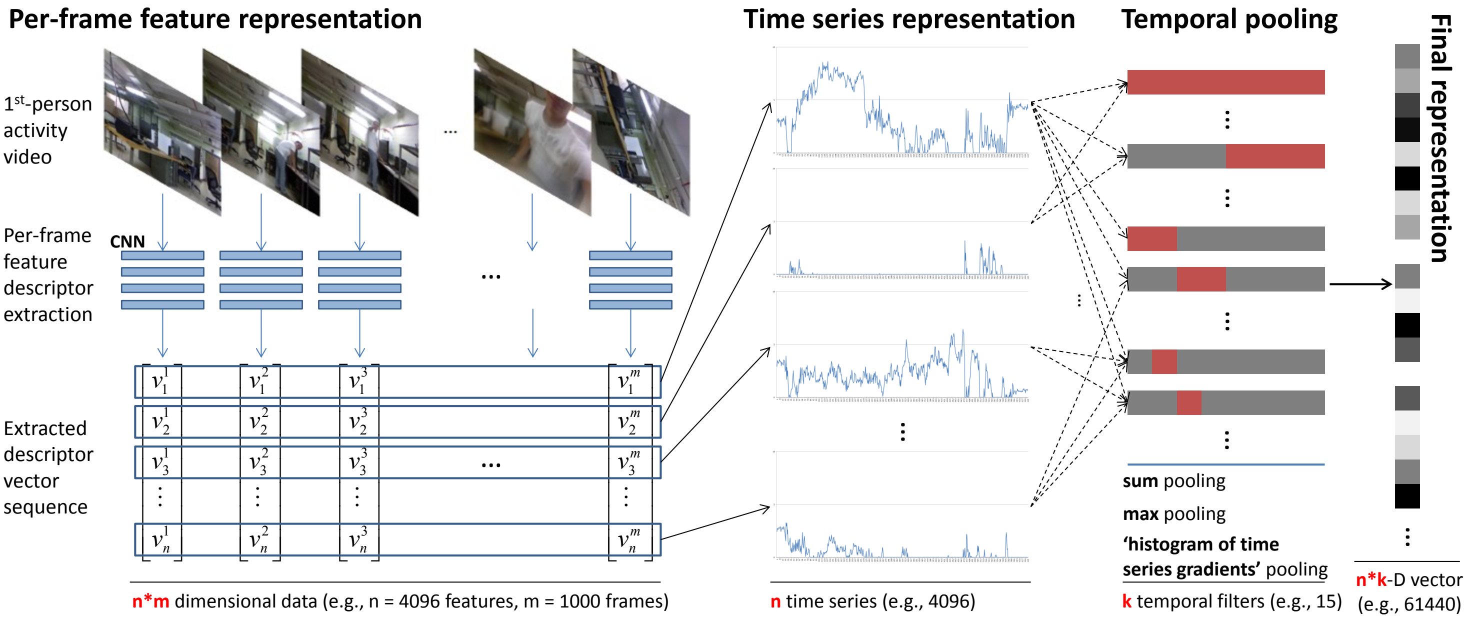
Per-frame feature representation
1st-person activity video
Per-frame feature descriptor extraction
CNN
Extracted descriptor vector sequence
$v_1^1$ $v_1^2$ $v_1^3$ $v_1^m$
$v_2^1$ $v_2^2$ $v_2^3$ $v_2^m$
$v_3^1$ $v_3^2$ $v_3^3$ $v_3^m$
$v_n^1$ $v_n^2$ $v_n^3$ $v_n^m$
n*m dimensional data (e.g., n = 4096 features, m = 1000 frames)
Time series representation
n time series (e.g., 4096)
Temporal pooling
sum pooling
max pooling
'histogram of time series gradients' pooling
k temporal filters (e.g., 15)
Final representation
n*k-D vector (e.g., 61440)

Figure 1: Overall representation framework of our pooled time series (PoT). Given a sequence of per-frame feature descriptors (e.g., HOF or CNN features) from a video, PoT represents motion information in the video by computing short-term/long-term changes in each descriptor value.

In this paper, we present a new feature representation for first-person videos. In first-person video understanding (e.g., activity recognition [4]), it is very important to capture both entire scene dynamics (i.e., egomotion) and salient local motion observed in videos. We describe a representation framework based on time series pooling, which is designed to abstract short-term/long-term changes in feature descriptor elements. The idea is to keep track of how descriptor values are changing over time and summarize them to represent motion in the activity video. The framework is general, handling any types of per-frame feature descriptors including conventional motion descriptors like histogram of optical flows (HOF) as well as appearance descriptors from more recent convolutional neural network (CNN). We name our representation as *pooled time series* (PoT), and Figure 1 shows its overall framework.

We experimentally confirm that our approach clearly outperforms previous feature representations including bag-of-visual-words (BoW) and improved Fisher vector (IFV) [3] when using identical underlying feature descriptors. Multiple first-person activity datasets [1, 2] were tested under various settings to confirm these findings. Figure 2 compares performances of our PoT representations with BoW and IFV while using four types of feature descriptors (i.e., HOF, MBH, Overfeat, and Caffe) and their combinations. In addition, we also confirmed that our PoT representation has superior performance to existing state-of-the-art features like local spatio-temporal features and Improved Trajectory Features [5] (originally developed for 3rd-person videos) when handling first-person videos. Even with the temporal pyramid pooling added to the original ITF, our PoT performed much better than the ITF: 0.676 vs. 0.730 for the DogCentric Activity dataset, and 0.766 vs. 0.794 for the UEC-Park dataset.

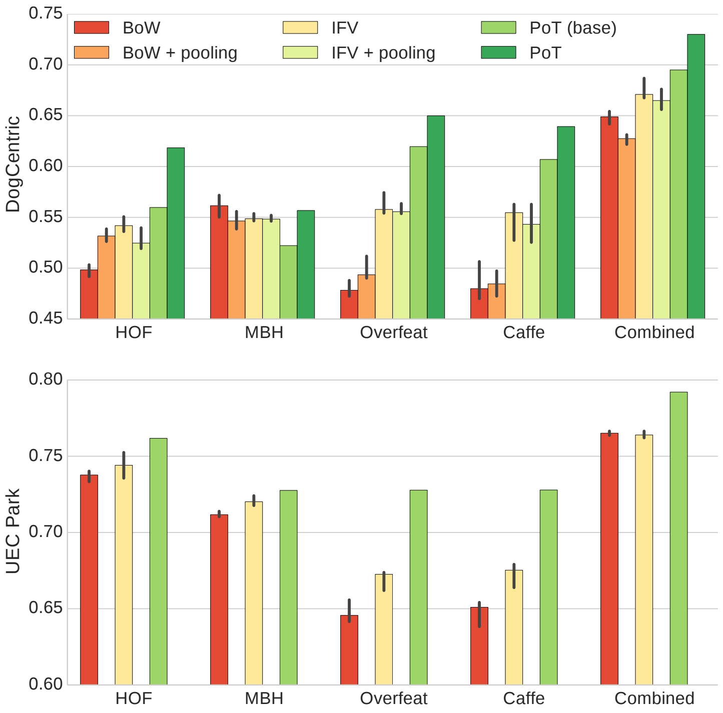

Figure 2: Classification accuracies of feature representations with each descriptor (and their combination). Representations that utilize randomness are drawn with 95% confidence intervals. See the paper for details.

[1] Y. Iwashita, A. Takamine, R. Kurazume, and M. S. Ryoo. First-person animal activity recognition from egocentric videos. In *ICPR*, 2014.

[2] Kris M. Kitani, Takahiro Okabe, Yoichi Sato, and Akihiro Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *CVPR*, 2011.

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

[4] M. S. Ryoo and L. Matthies. First-person activity recognition: What are they doing to me? In *CVPR*, 2013.

[5] Heng Wang and Cordelia 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.