Pooled Motion Features for First-Person Videos

M. S. Ryoo, Brandon Rothrock, and Larry Matthies
Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA
mryoo@jpl.nasa.gov

Abstract

In this paper, we present a new feature representation for first-person videos. In first-person video understanding (e.g., activity recognition), 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 networks (CNN).

We experimentally confirm that our approach clearly outperforms previous feature representations including bag-of-visual-words and improved Fisher vector (IFV) when using identical underlying feature descriptors. We also confirm that our feature representation has superior performance to existing state-of-the-art features like local spatio-temporal features and Improved Trajectory Features (originally developed for 3rd-person videos) when handling first-person videos. Multiple first-person activity datasets were tested under various settings to confirm these findings.

1. Introduction

First-person videos, also called egocentric videos, are videos taken from an actor's own viewpoint. The volume of egocentric video is rapidly increasing due to the recent ubiquity of small wearable devices. The main difference between conventional 3rd-person videos and 1st-person videos is that, in 1st-person videos, the person wearing the camera is actively involved in the events being recorded. Strong egomotion is observed in first-person videos, which makes them visually very unique (Figure 1). Automated understanding of such videos (i.e., first-person activity recognition) is crucial for many societal applications including quality-of-life systems to support daily living and video-based life-logging. Applications also include robot perception and human-robot interactions, since videos from the robot's viewpoint naturally are in first-person.

Despite a massive amount of first-person videos becoming available, approaches to semantically understand such videos have been very limited. This is particularly true for research on 'motion features' for first-person videos, which serves as a fundamental component for visual grounding of actions and events. Even though there has been previous works on extraction of first-person-specific semantic features like hand locations [11] and human gaze [13], features and representations designed to capture motion dynamics of first-person videos have been lacking. Representing this egomotion is very essential for recognition of sports activities, accident activities for patient/health monitoring (e.g., a person collapsing), activities for surveillance/military (e.g., another person assaulting), and many others from first-person videos. Most of the previous first-person activity recognition works [9, 18] focused on the use of existing features and representations designed for conventional 3rd-person videos, without tailoring motion features for the first-person case.

This paper introduces a new feature representation named *pooled time series* (PoT). Our PoT is a general representation framework based on time series pooling of feature descriptors, which is particularly designed to capture motion information in first-person videos. Given a sequence of per-frame feature descriptors (e.g., HOF or CNN features) from a video, PoT abstracts them by computing short-term/long-term changes in each descriptor element. The motivation is to develop a new feature representation that captures 'details' of entire scene dynamics displayed in first-person videos, thereby obtaining better video recognition performances. Capturing egomotion information is crucial for recognition of ego-actions and interactions from first-person videos, and our PoT representation allows the system to do so by keeping track of very detailed changes in feature descriptor values while suppressing noise. Multiple novel pooling operators are introduced, and are combined with temporal filters to handle the temporal structure of human activities.

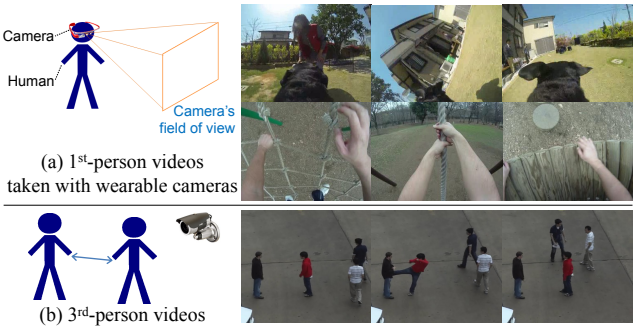


Figure 1. Conceptual comparison between 1st-person videos and 3rd-person videos. Example snapshots from public first-person datasets [9, 5] taken with human/animal wearable cameras and those from public 3rd-person dataset [17] are also illustrated.

We experimentally confirm that our proposed PoT representation clearly outperforms previous feature representations such as bag-of-visual-words and improved Fisher vector [14] on first-person activity recognition. Both the global motion aspect and local motion aspect of first-person videos are captured with our PoT by taking advantage of different types of descriptors, and we illustrate recognition accuracies of our PoT with each of the descriptors as well as their combinations. Furthermore, we demonstrate that our combined PoT representation has superior performance to the best-known motion feature designed for 3rd-person videos [21], when handing 1st-person videos.

1.1. Related works

Recognition from first-person videos is a topic with an increasing amount of attention. There are works focusing on first-person-specific features, including hand locations in first-person videos [11] and human gaze estimation based on first-person videos [13]. There also have been works on object recognition from first-person videos [12, 15].

However, study on motion features for first-person videos has been relatively limited, particularly those for first-person activity recognition. Most of the works focused on temporal segmentation of videos using optical flow-based features, without taking advantage of high-dimensional image features for detailed recognition of high-level activities. Kitani *et al.* [9] worked on unsupervised learning of ego-actions and segmentation of videos based on it. A simple histogram based on optical flow direction/magnitude and frequency was constructed as a feature representation, which can be viewed as an extension of HOF. Poleg *et al.* [16] introduced the use of displacement vectors similar to optical flows for long-term temporal segmentation, but they only focused on segmentation of relatively simple egomotion such as walking and wheeling. [18] investigated the first-person activity recognition scenarios by combining multiple features, while particularly focusing on recognition of interaction-level activities. Still, they used very conventional HOF and local spatio-temporal features [10, 3] together with general bag-of-visual-words representation, without any attempt to develop first-person-specific features.

2. Pooled times series representation

In this section, we introduce our new feature representation named *pooled time series* (PoT), which is specifically designed for first-person videos. The role of a feature 'representation' is to abstract a set of raw feature descriptors (e.g., histogram of oriented gradients) extracted from each video into a single vector representing the video. It converts a large number of high-dimensional descriptors into a single vector with a tractable dimensionality, allowing its result to serves as an input vector for classifiers (e.g., activity classification). Existing feature representations include bag-of-visual-words (BoW) and improved Fisher vector (IFV), which converts a set of raw descriptors into a low-dimensional histogram. What we introduce in this section is a new feature representation that better abstracts motion displayed in first-person videos.

The overall pipeline of our PoT representation is as follows. Given a first-person video (i.e., a sequence of image frames), our approach first extracts appearance/motion descriptors from each frame. As a result, a sequence of n-dimensional descriptor vectors is obtained where n is the size of the vector from each frame. Our approach interprets this as a set of n time series. The idea is to keep track of how each element of the descriptor vector is changing over time (i.e., it becomes a function of time), and summarize such information to represent the activity video. Next, temporal pooling is performed: a set of temporal filters (i.e., time intervals) is applied to each time series and the system performs multiple types of pooling operations (e.g., max, sum, gradients, ...) per filter. Finally, the pooling results are concatenated to form the final representation of the video. Figure 2 illustrates the overall process.

Let each per-frame feature descriptor obtained at frame t be denoted as $V^t = [v_1^t, v_2^t, ..., v_n^t]$. Our PoT representation framework interprets this sequence of vectors $V^1, ..., V^m$ (m is the number of video frames) as a set of time series, $\{f_1(t), ..., f_n(t)\}$. That is, each of our time series $f_i(t)$ corresponding to the ith feature descriptor value is defined as $f_i(t) = v_i^t$. For each time series, temporal pooling is performed with a set of k temporal filters, which essentially is a set of time intervals to make the system focus on each local time window: $\{[t_1^s, t_1^e], ..., [t_k^s, t_k^e]\}$. A temporal pyramid structure [2] is used in our implementation to obtain filters, but any number of filters with (overlapping) intervals can be used by our framework in principle.

Finally, multiple pooling operators are applied for each filter and their results are concatenated to obtain the final

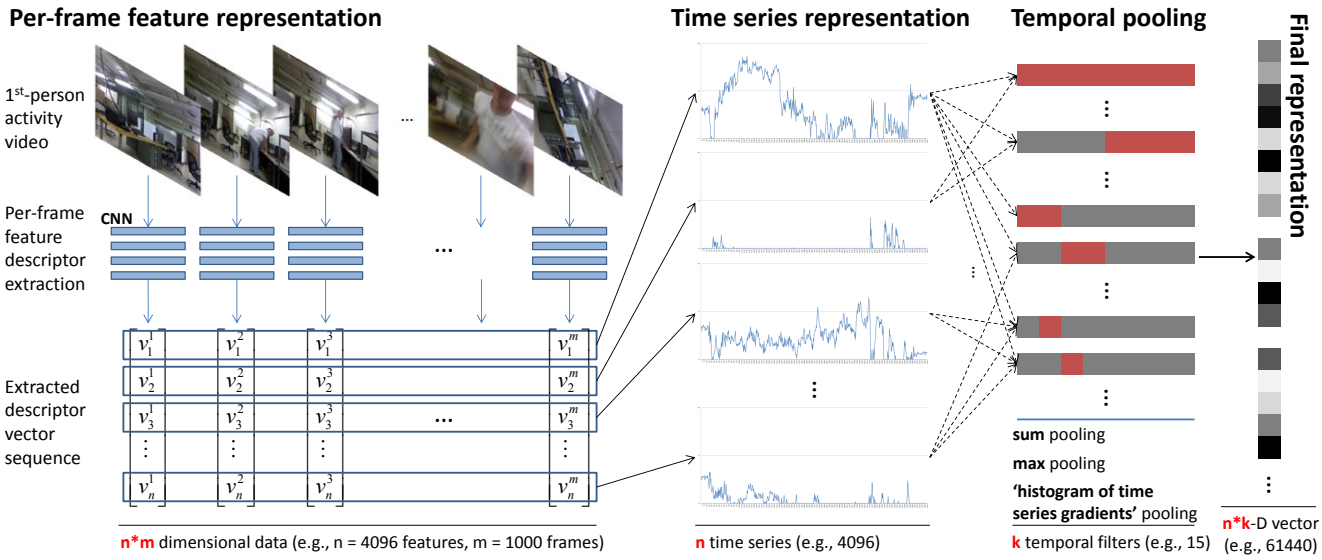


Figure 2. Overall representation framework of our pooled time series (PoT).

PoT feature representation of the video:

$$x = [x_1^{op_1}[t_1^s, t_1^e],\ x_1^{op_2}[t_1^s, t_1^e], \cdots,\ x_n^{op_r}[t_k^s, t_k^e]] \quad (1)$$

where $x_i^{op_j}$ specifies that it is applying the jth pooling operator to the ith time series $f_i(t)$. Our PoT representation takes advantage of four different types of pooling operators including two newly introduced temporal pooling operators, which we discuss more in Subsection 2.2.

Our framework of (i) extracting per-frame descriptors, (ii) interpreting them as a set of time series, and (iii) applying various types of time series pooling and concatenating them provides the following three important abilities:

First, (1) it preserves detailed dynamics displayed in each descriptor element as a time series, and allows the representation to capture both long-term motion and short-term information with multiple temporal filters. That is, depending on the nature of the time series, our representation is able to capture subtle short-term motion by pooling from a filter with a small time interval as well as long-term motion by performing pooling with a large time interval. Such flexibility is in contrast to previous bag-of-visual-words representation for global motion descriptors (e.g., the one used in [18]) that abstracts all descriptor values in one frame (or a subsequence of few frames) into a single discretized 'visual word'. In addition, (2) our representation explicitly imposes temporal structure of the activity by decomposing the entire time interval to multiple subintervals, which is very important for representing high-level activities. Finally, (3) it allows the system to take advantage of multiple types of pooling operators so that the representation captures different aspects of the data.

As a result of our framework, each video is represented with one single vector having a tractable dimensionality. Activity recognition is performed by training/testing standard classifiers (e.g., SVM classifiers) with these vectors. Our representation is able to cope with any type of generative and discriminative classifiers in principle, and we show its superiority over others in Section 3.2.

2.1. Handling high-dimensional feature descriptors

The proposed representation framework is very general in the aspect that it is able to cope with any types of per-frame image/motion descriptors such as histogram of oriented gradients (HOG) or histogram of optical flows (HOF). Furthermore, it is particularly designed to handle high-dimensional per-frame image descriptors: image-based deep learning features which are also called convolutional neural network (CNN) features [7, 19]. These deep learning image features are obtained by snatching intermediate outputs from internal convolutional layers of a CNN, pre-trained on image datasets. They can also be viewed as cascades of automatically learned image filters. These image descriptors are trained from large scale image datasets and have obtained highly successful results on image classification/detection [4] as well as video classification [8], performing superior to state-of-the-art hand-designed image descriptors such as HOG even without re-training the networks.

Our motivation was to design a general representation that best takes advantage of such high-dimensional descriptors and confirm that these CNN features are able to increase first-person activity recognition performance significantly together with other features. Each element of a CNN feature vector abstracts particular local/global appearance for a single frame, and (by extension) its time series models how this local/global appearance is changing

over time. As a human in the scene moves (e.g., changes his/her posture) and the camera changes its viewpoint because of egomotion, certain CNN feature values will become activated/deactivated and our idea is to keep track of such changes to represent the activity video. In our experiments, we explicitly confirm this while comparing our representation with the conventional representations. When using CNN features as our base per-frame descriptors, we get a 4096-dimensional feature vector (i.e., n=4096) for each image frame by obtaining outputs of the last convolutional layer (e.g., stage 7 in [19]).

2.2. Temporal pooling operators

Our PoT representation is constructed by applying multiple types of temporal pooling operators over each temporal filter (i.e., time interval). In this paper, we take advantage of four different types of pooling operators: conventional max pooling and sum pooling, and two types of our new 'histogram of time series gradients' pooling.

Our max pooling and sum pooling operators are defined as follows:

$$x_i^{\max}[t^s, t^e] = \max_{t=t_s..t_e} f_i(t),$$
$$x_i^{\sum}[t^s, t^e] = \sum_{t=t_s}^{t_e} f_i(t). \tag{2}$$

In addition to these traditional pooling operators, we newly introduce the concept of 'histogram of time series gradients' pooling. The idea is to count the number of positive (and negative) gradients within the temporal filter.

$$x_i^{\Delta_1^+}[t^s, t^e] = |\{t \mid f_i(t) - f_i(t-1) > 0 \wedge t^s \le t \le t^e\}|,$$
$$x_i^{\Delta_1^-}[t^s, t^e] = |\{t \mid f_i(t) - f_i(t-1) < 0 \wedge t^s \le t \le t^e\}|. \tag{3}$$

Furthermore, we propose a variation of the above new pooling operator, which sums the amount of positive (or negative) gradients instead of simply counting their numbers. It is defined as:

$$x_i^{\Delta_2^+}[t_s, t_e] = \sum_{t=t_s}^{t_e} h_i^+(t), \quad x_i^{\Delta_2^-}[t_s, t_e] = \sum_{t=t_s}^{t_e} h_i^-(t) \tag{4}$$

where

$$h_i^+(t) = \begin{cases} f_i(t) - f_i(t-1) & \text{if } (f_i(t) - f_i(t-1)) > 0 \\ 0 & \text{otherwise,} \end{cases}$$
$$h_i^-(t) = \begin{cases} f_i(t-1) - f_i(t) & \text{if } (f_i(t) - f_i(t-1)) < 0 \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

Each of our time series gradients pooling operation generates a pair of values (i.e., $x_i^{\Delta^+}$ and $x_i^{\Delta^-}$) instead of a single value like max pooling. These two values are concatenated in our PoT representations.

3. Experiments

3.1. Experimental settings

Datasets: We conducted our experiments with two different public first-person video datasets: DogCentric activity dataset [5] and UEC Park dataset [9]. These are very challenging datasets with strong camera egomotion, which are different from conventional 3rd-person datasets. Figure 1 shows sample images. DogCentric dataset was recorded with wearable cameras mounted on dogs' back. UEC Park dataset was collected by a human wearing a camera. DogCentric dataset contains ego-actions of the dog as well as interactions between the dog and other humans (e.g., a human throws a ball and the dog chases it). UEC Park dataset contains ego-actions of the person (wearing a camera) involved in various types of physical activities (e.g., climbing a ladder) at a park. Dogcentric dataset consists of 10 activity classes, while UEC dataset consists of 29 classes.

Representation implementation: We implemented our PoT representations with four different types of per-frame feature descriptors: histogram of optical flows (HOF), motion boundary histogram (MBH) used in [21], Overfeat CNN image feature [19], and Caffe CNN image feature [7]. The first two descriptors (i.e., HOF and MBH) are optical flow based motion descriptors, and the last two (i.e., Overfeat and Caffe) are deep learning based image appearance descriptors from CNNs pre-trained on ImageNet. Our HOF descriptors are in 200-D (5-by-5-by-8), MBH descriptors are in 400-D (two 5-by-5-by-8), and Overfeat and Caffe are in 4096-D. L1 normalization was applied for each descriptor. As a result, four different versions of our PoT representations were implemented as well as the final representation combining all four representations. As described in Section 2, pyramid temporal filters with level 4 were used and four different types of pooling operators were applied.

Classifiers: In all our experiments, we used the same nonlinear SVM with a χ^2 kernel. It showed better performance compared to linear SVM. When combining representations with multiple descriptors, a multi-channel kernel was used.

Evaluation setting: We followed the standard evaluation setting of the DogCentric dataset: we performed repeated random training/testing splits 100 times, and averaged the performance. We randomly selected half of videos per activity class as training videos, and used the others for the testing. If the number of total videos per class is odd, we made the testing set to contain one more video. Once training videos are selected, they are used across the entire experiments for fair comparisons.

3.2. Feature representation evaluation

We conducted experiments to confirm superiority of our proposed PoT representation over conventional feature rep-

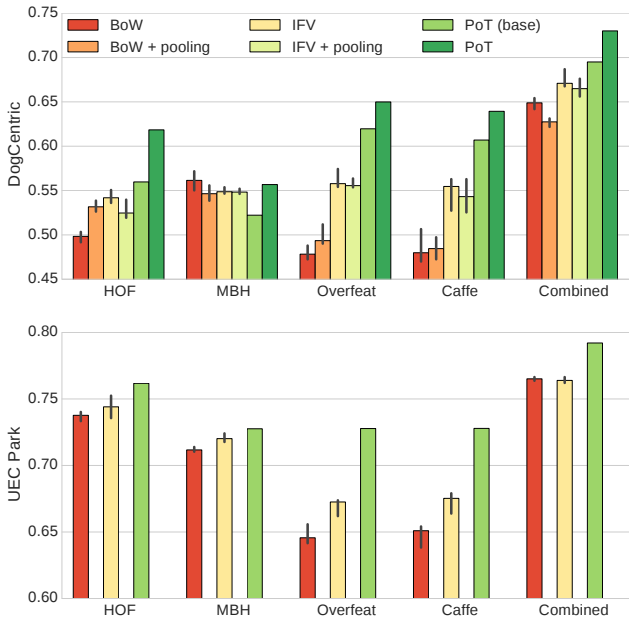


Figure 3. Classification accuracies of feature representations with each descriptor (and their combination). Representations that utilize randomness are drawn with 95% confidence intervals. See text for details.

resentations. The idea is to evaluate the performances of our PoT, and compare it with those of other widely-used state-of-the-art representations while making them use exactly the same feature descriptors. More specifically, we compared our PoT representations with bag-of-visual-words (BoW) and Improved Fisher Vector (IFV) [14] while using four types of feature descriptors (i.e., HOF, MBH, Overfeat, and Caffe) and their combinations.

BoW and IFV are very commonly used feature representations in the activity recognition literature [1]. BoW represents a video as a histogram of 'visual words' by clustering all feature descriptors, making each descriptor assigned to a specific visual word. IFV can be viewed as a 'soft' version of that, in the aspect that it represents each descriptor as a set of soft assignments to cluster centers. We tested tens of different parameter settings for each feature type, and chose the best setting per feature. This includes the tuning of the number of visual words (e.g., IFV has 4000-D for HOF and 40960-D for Caffe). In addition, we implemented BoW and IFV in conjunction with the temporal pyramid pooling identical to the one used in our PoT, so that they also consider temporal structure among features. We explicitly compared all these different representations with our PoT. Also, since the clustering processes of BoW and IFV contain randomness, we report their 95% confidence interval together with the median performance by testing them 10 times.

DogCentric activity dataset: Figure 3 (top) describes the 10-class activity classification accuracies of our representation (and BoW and IFV) for each of the base descriptors.

Here, we are showing the accuracies of the PoT representation with the best combination of pooling operators. As described in Section 2.2, there are four different pooling operators our PoT representation can take advantage of. We conducted experiments with all possible combinations of pooling operators for PoT (which can be found in our supplementary Appendix), and selected the best performing combination. In general, concatenations of all pooling operators (e.g., $\sum$+max+Δ_1) obtained the best results, or results very close to the best. 'PoT (base)' is the basic version of our feature representation, which is constructed by applying the pooling operator to a single time interval that covers the entire activity video (i.e, no temporal pyramid structure).

We are able to observe that our PoT representations perform superior to BoW and IFV in all cases, except for MBH descriptors where all representations showed similar performances. Even when we add temporal pyramid pooling (identical to the one used in our representation) to BoW and IFV, their performances were clearly inferior to our PoT. The mean accuracy of the combined IFV representation (with pyramid) was 0.666, while our PoT obtained the accuracy of 0.730. Previous state-of-the-art is 0.605 [5].

Particularly, in the case of using high-dimensional deep learning features (i.e., 4096-D in Overfeat and Caffe), we confirmed that our representation significantly improves the performance over both BoW and IFV. We believe this is due to the fact that per-frame abstraction (i.e., clustering) performed in BoW and IFV fails to capture subtle local cues while our representation is particularly designed to handle such high-dimensionality descriptors. BoW abstracts each high-dimensional per-frame descriptor into a single 'visual word' (or a few soft assignments in case of IFV). As a result, subtle changes in a small number of descriptor values are ignored in general, which is particularly harmful in the case of high-dimensional descriptors. This is unlike our PoT that tries to explicitly capture such changes with time series pooling. The result suggests that PoT is the better representation to take advantage of modern high-dimensional feature descriptors.

Another important observation is that our PoT benefitted greatly by considering temporal structures among features, much more compared to BoW and IFV. We discuss this more in Subsection 3.3.

In addition, since dynamic time warping (DTW) is a traditional approach to deal with time series data, we also tested a basic DTW-based template matching (using the same time series with PoT) as a baseline. The best performance of DTW was 0.288, as opposed to 0.730 of ours.

UEC Park dataset: We also performed the same experiments described above with one more public first-person video dataset: UEC Park dataset. As described in the previous subsection, the dataset contains video segments labeled

with 29 different classes. Labels are very rough and the number of videos per activity class are very unbalanced in this dataset (e.g., there is a class with only 1 video), making the classification task challenging.

Also, videos of this dataset were obtained by segmenting a long video every 2 seconds, and each segment was labeled based on the most dominant activity observed in the segment. As a result, activities in these videos are not temporally aligned (e.g., a video segment may not even contain the initial part of the activity at all) and using pooling with temporal structures only harms the recognition performances. We confirmed this with all representations: BoW, IFV, and PoT. Thus, here, we only show the results of representations without any temporal structure consideration. PoT is at a disadvantage for this dataset, since it benefits greatly using pooling with temporal structures while BoW and IFV do not, as we discuss more in Subsection 3.3.

Figure 3 (bottom) shows the result. Our PoT obtained the best performance on all feature descriptors, similar to the case of the DogCentric dataset. PoT obtained the best result using all four feature descriptors and obtained particularly higher performances for high-dimensional CNN features. Even though PoT was not able to fully take advantage of activites' temporal structures, it still performed superior to BoW and IFV.

3.3. Temporal structure evaluation

We conducted further experiments to confirm the advantage of PoT: it benefits more compared to other representations when considering temporal structure among features. DogCentric activity dataset was used for this experiment. We illustrate classification accuracies of BoW, IFV, and PoT with and without consideration of the temporal pyramid structure. 'Without pyramid' means that the feature representation is constructed by applying the pooling operator on one single time interval that covers the entire activity video. 'With pyramid' means that a set of temporal filters were used. For PoT, we also compare results of different time series pooling operators (and their combinations) with and without temporal pyramids.

Figure 4 shows the results. We are able to confirm that consideration of temporal structure benefited the recognition with our PoT while it did not benefit the other representations much. This is particularly true for the representations with combinations of all four descriptors. We believe this observation is caused by the following characteristic: as mentioned in Subsection 3.2, abstraction/discretization of per-frame observations in BoW (or IFV) completely ignores subtle local descriptor changes. This makes them have less chance to capture short-term local motion even when we temporally split the video using a pyramid. On the other hand, PoT does not suffer from such abstraction.

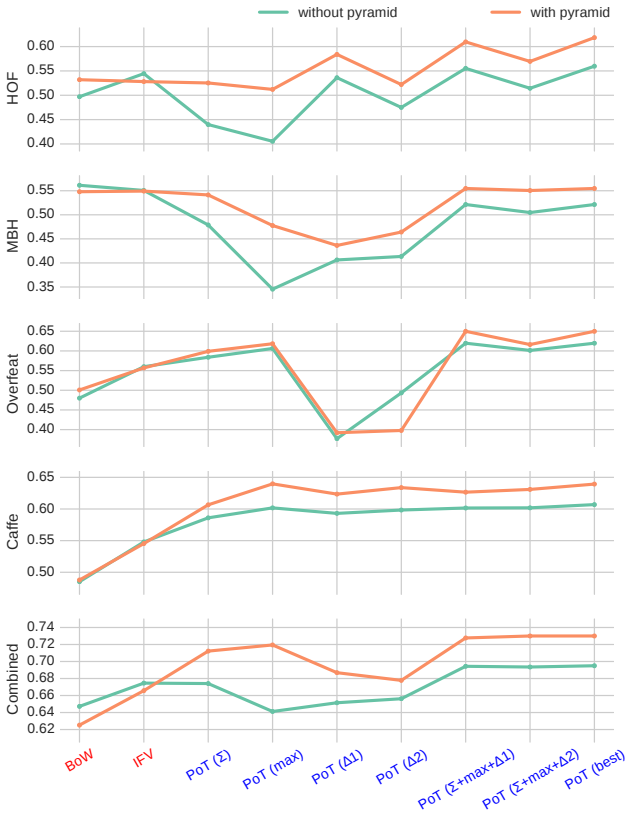


Figure 4. Feature performance using BoW and IFV compared with various PoT pooling operators with and without a temporal pyramid on the DogCentric dataset. Y-axis is classification accuracy, and X-axis shows different representations. PoT generally benefits much more from the temporal pyramid than BoW and IFV.

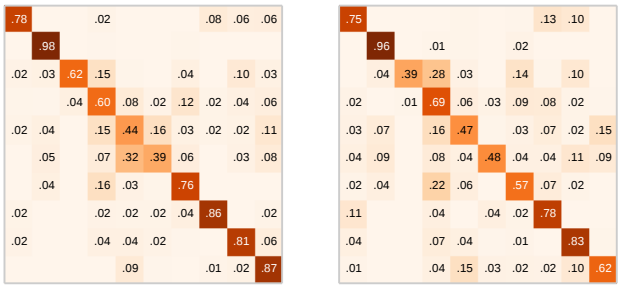
3.4. Comparison to state-of-the-art features

We explicitly compared activity classification accuracies of our PoT representations with other state-of-the-art features, including well-known local spatio-temporal features [10, 3] and recent trajectory-based local video features [21]. Notably, INRIA's improved trajectory feature (ITF) [21] is the one that obtained the best performance in the ICCV 2013 challenge on UCF101 dataset [20]. ITF internally takes advantage of three different feature descriptors similar to ours: HOG, HOF, and MBH. IFV representations were used for all these features. For the DogCentric dataset experiments, the temporal pyramid structure was considered by all of these previous features, since using temporal pyramid improved their classification accuracies by 0.03∼0.05. What we show is that our new feature representations (together with frame-based descriptors) are more suitable for representing motion in first-person videos compared to the previous features designed for 3rd-person videos.

In addition, we implemented the approach of combining ITF with CNN descriptors [6], which won the ECCV 2014 classification challenge on UCF101 dataset. Both Overfeat

Table 1. A table comparing performances of the proposed approach with state-of-the-arts on DogCentric dataset [5]: F1-scores per class and the final 10-class classification accuracies. Approaches with our representations are colored blue. The performances are split into three categories: representations with only one descriptor, representations with multiple descriptors, and combinations of multiple different features representations. The best performance per category is indicated with bold. The overall best performance is indicated with bold+red.

	Ball play	Waiting car	Drink	Feed	Turn head (left)	Turn head (right)	Pet	Body shake	Sniff	Walk	Final accuracy
Single descriptor features											
STIP (with IFV) [10]	0.579599	0.767359	0.453599	0.451974	0.407274	0.327223	0.425169	0.788209	0.657433	**0.695971**	0.5764537 ± 0.008
Cuboid (with IFV) [3]	0.646436	0.777357	0.535933	0.471589	0.252507	0.068375	0.550717	0.854224	0.720194	0.682976	0.5961668 ± 0.007
IFV - HOF	0.613862	0.619561	0.421567	0.308094	0.345337	0.281042	0.548307	0.747027	0.535438	0.662877	0.5281018 ± 0.01
IFV - MBH	0.641111	0.863174	0.261308	0.470158	0.332133	0.284172	0.443836	0.674957	0.61996	0.570858	0.549259 ± 0.006
IFV - Overfeat	0.614632	0.815432	0.452256	0.576355	0.311339	0.318194	0.676456	0.465434	0.642253	0.449071	0.5567592 ± 0.004
IFV - Caffe	0.660997	0.860932	0.58083	0.55149	0.269409	0.229007	0.639273	0.446267	0.649675	0.411165	0.5453332 ± 0.014
PoT - HOF (Σ + Δ1)	**0.790159**	0.79585	0.586986	0.431446	**0.499128**	0.468511	0.501162	0.832549	0.590227	0.671908	0.618426
PoT - MBH (Σ + max + Δ1)	0.644583	0.709702	0.320233	0.386772	0.441355	**0.567058**	0.299373	**0.873127**	0.574558	0.627518	0.556759
PoT - Overfeat (Σ + max + Δ1)	0.74655	0.895397	0.640212	0.594052	0.291462	0.355681	**0.783395**	0.726989	0.755053	0.564199	**0.649907**
PoT - Caffe (max + Δ2)	0.738111	**0.900798**	**0.725664**	**0.626889**	0.33498	0.338351	0.705605	0.592798	**0.773768**	0.545347	0.639352
Multi-descriptor features											
Inria ITF (with IFV) [21]	0.691291	0.893157	0.545962	0.579966	**0.495152**	**0.589468**	0.625639	0.708337	0.778854	0.676454	0.6757592 ± 0.006
IFV - all	0.753374	0.876573	0.580573	**0.597431**	0.368204	0.305966	0.725102	0.789129	0.742027	0.659151	0.6657036 ± 0.008
PoT - all (Σ + max + Δ1)	**0.820552**	**0.932507**	0.68982	0.59662	0.45	0.472542	**0.758327**	0.854455	0.817615	0.778485	0.727685
PoT - all (Σ + max + Δ2)	0.820359	0.93047	**0.714618**	0.584604	0.439639	0.45971	0.756757	**0.870967**	**0.82741**	**0.788683**	**0.73**
Combinations of multiple feature representations											
Iwashita et al. 2014 [5]	0.618939	0.818613	0.383081	0.510749	0.397806	0.41918	0.544725	0.86418	0.698887	0.779148	0.605
STIP + Cuboid (with IFV)	0.685341	0.788416	0.471734	0.519026	0.395602	0.23537	0.540914	0.837125	0.74858	0.738611	0.6291759 ± 0.008
ITF + STIP + Cuboid	0.714646	0.871098	0.533088	0.591699	0.47348	0.423435	0.650097	0.827838	0.813176	0.753727	0.6912039 ± 0.006
ITF + CNN [6]	0.696641	0.928735	0.703593	**0.651387**	0.434422	0.415808	**0.778585**	0.726934	0.808843	0.608596	0.692315 ± 0.004
PoT + STIP + Cuboid	0.804031	0.925272	0.712457	0.591944	0.460633	0.433242	0.742753	0.866876	0.836603	0.797134	0.73137 ± 0.001
PoT + ITF	**0.826126**	**0.933284**	**0.71304**	0.597523	0.477482	**0.515245**	0.754544	**0.87818**	**0.851537**	0.795619	**0.7447038 ± 0.001**
PoT + ITF + STIP + Cuboid	0.819623	0.92363	0.703833	0.594038	**0.479739**	0.50065	0.733664	0.873437	0.848916	**0.809405**	0.7406666 ± 0.001



PoT (Combined) INRIA ITF

Figure 5. Confusion matricies on the DogCentric dataset, comparing our PoT (combined) and the state-of-art INRIA ITF feature (with IFV).

and Caffe were used, and mean of per-frame CNN vectors were computed and added to ITF.

Table 1 shows the results with the DogCentric dataset. In addition to the final 10-class classification accuracies of the approaches, we are also reporting per-class F1-scores of them. The motivation is to analyze which feature descriptor/representation is better for recognition of which activity class. Instead of simply reporting per-class classification accuracies that do not take false positives into account, we computed a pair of precision and recall values per class from the confusion matrix and obtained F1-scores based on them.

The result clearly illustrate that our PoT obtains the best result, outperforming local spatio-temporal features as well

Table 2. A table comparing performances of the proposed approach with state-of-the-arts using UEC Park dataset [9]: 29-class classification accuracies are shown.

	Final accuracy
STIP (with IFV) [10]	0.6913438 ± 0.003
Cuboid (with IFV) [3]	0.7233332 ± 0.002
BoW - all	0.7649616 ± 0.002
IFV - all	0.7640002 ± 0.002
Inria ITF (with IFV) [21]	0.7662412 ± 0.002
ITF + CNN [6]	0.757359 ± 0.002
PoT - all (best)	**0.793641**
PoT + ITF	**0.794897 ± 0**

as the previously reported results [5]. Particularly, our PoT performed significantly superior to the state-of-the-art ITF approach. Even with the temporal pyramid pooling added to the original ITF, our PoT performed much better than the ITF: 0.676 vs. 0.730. The ITF performance without pyramid was 0.638. Our PoT also showed the best per-class recognition accuracies in most of the classes. ITF showed slightly better performances for 'turn head' classes, since these videos are actually more similar to 3rd-person videos: the camera was mounted on the back of the dog (i.e., not head) and these 'turn head' videos do not involve much camera motion unlike the others. Furthermore, our PoT performed superior to the conventional method [6] of combining ITF and mean per-frame CNN: 0.692 vs. 0.730.

We are able to observe similar results with the UEC Park dataset. Table 2 shows the results. PoT obtained the best

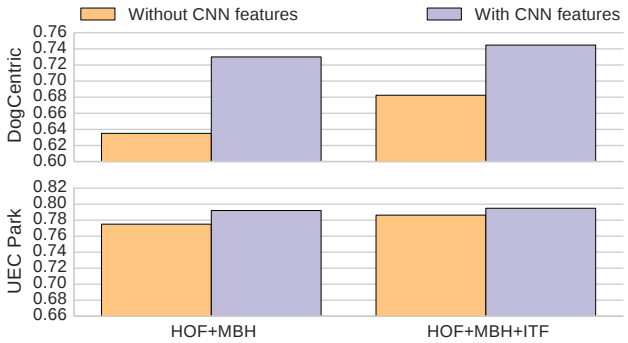


Figure 6. Performance gain from combining CNN features with conventional motion features for both datasets.

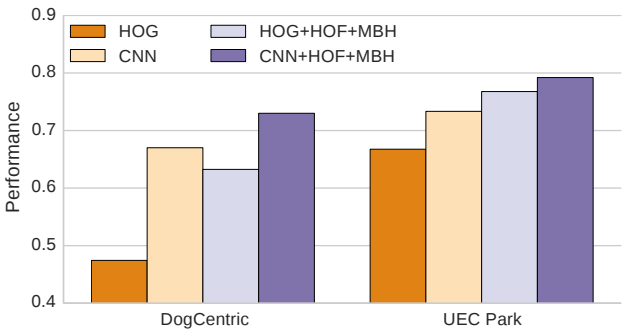


Figure 7. Performance comparison of CNN features as a replacement for HOG, showing consistent gains on both datasets with and without motion features included.

performance, and we were able to (slightly) increase the performance further by combining PoT with ITF.

3.5. Evaluation of appearance-based features

Taking advantage of CNN descriptors: We explicitly compared the recognition accuracies of our approach with and without CNN-based appearance descriptors. The idea was to confirm 'how much benefit our PoT representation is able to get from CNN descriptors' when representing first-person videos for activity recognition. The result is that, with our PoT, CNN-based appearance descriptors capture information different from motion descriptors and combining them with others really benefits the entire system, while the degree of their effectiveness is dependent on the dataset and activities. Figure 6 shows the results.

For the DogCentric dataset, using CNN descriptors greatly benefited the overall recognition accuracy. Notice that CNN descriptors themselves showed superior performance compared to all other motion-based descriptors (e.g., HOF) with our PoT, as described in Figure 3. DogCentric dataset contains activity videos taken at various environments (indoor, outdoor, ...), and certain activities are highly correlated with such environment/background information (e.g., there will not be 'ball chasing' activity in an indoor environment). As a consequence, capturing appearance information is very important for these activities/videos, and CNN descriptors showed very good results on them with our PoT. On the other hand, all UEC Park video sequences are taken at a same environment (i.e., a park), and thus CNN-based appearance descriptors were not as effective as motion descriptors. Nevertheless, in both cases, combining CNN features with other descriptors benefited the overall recognition performances, suggesting that our PoT is properly taking advantage of them.

Appearance descriptors: CNN vs. HOG: We tested another appearance descriptor, histogram of oriented gradients (HOG), and compared it with the CNN descriptors we are using. For this experiment, we extracted pure histogram of oriented gradients similar to our HOF from images. For each frame, a HOG descriptor was constructed with 8 different gradient directions and 5-by-5 spatial bins. Then, each sequence of these HOG descriptors was represented using our PoT representation. We not only compared our PoT representation only based on HOG with those only based on CNN descriptors, but also tested our final 'combined' PoT representing using both appearance descriptors and motion descriptors (i.e., HOF and MBH) by replacing CNN with HOG.

The idea was to compare two appearance descriptors (CNN vs. HOG) in representing first-person videos, and confirm that our PoT is appropriately taking advantage of CNN descriptors which is supposed to perform superior to HOG. Figure 7 illustrates the results obtained with the two datasets we use. We are able to observe that, with our PoT, CNN descriptors are significantly outperforming HOG descriptors under identical settings.

4. Conclusion

We introduced the new feature representation designed for first-person videos: *pooled time series* (PoT). Our PoT was designed to capture entire scene dynamics as well as local motion in first-person videos by representing long-term/short-term changes in high-dimensional feature descriptors, and it was combined with four different types of per-frame descriptors including CNN features. We evaluated our PoT using two public first-person video datasets, and confirmed that our PoT clearly outperforms previous feature representations (i.e., BoW and IFV) as well as the other state-of-the-art video features.

Acknowledgement: The research described in this paper was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. This research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-10-2-0016.

References

[1] J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review. *ACM Computing Surveys*, 43:16:1–16:43, April 2011.

[2] J. Choi, W. Jeon, and S. Lee. Spatio-temporal pyramid matching for sports videos. In *ACM MIR*, 2008.

[3] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *IEEE Workshop on VS-PETS*, 2005.

[4] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *arXiv:1311.2524*, 2013.

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

[6] M. Jain, J. van Gemert, and C. G. M. Snoek. University of amsterdam at thumos challenge 2014. In *ECCVW*, 2014.

[7] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv:1408.5093*, 2014.

[8] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

[9] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *CVPR*, 2011.

[10] I. Laptev. On space-time interest points. *IJCV*, 64(2):107–123, 2005.

[11] S. Lee, S. Bambach, D. J. Crandall, J. M. Franchak, and C. Yu. This hand is my hand: A probabilistic approach to hand disambiguation in egocentric video. In *CVPRW*, 2014.

[12] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, 2012.

[13] Y. Li, A. Fathi, and J. M. Rehg. Learning to predict gaze in egocentric video. In *ICCV*, 2013.

[14] F. Perronnin, J. Sanchez, and T. Mensink. Improving the Fisher kernel for large-scale image classification. In *ECCV*, 2010.

[15] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *CVPR*, 2012.

[16] Y. Poleg, C. Arora, and S. Peleg. Temporal segmentation of egocentric videos. In *CVPR*, 2014.

[17] M. S. Ryoo and J. K. Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *ICCV*, 2009.

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

[19] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *ICLR*, 2014.

[20] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human action classes from videos in the wild. *CRCV-TR-12-01*, 2012.

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