Modeling Sub-Event Dynamics in First-Person Action Recognition

Hasan F. M. Zaki[†*], Faisal Shafait[‡] and Ajmal Mian[†]
[†] School of Computer Science and Software Engineering, The University of Western Australia
[‡] National University of Sciences and Technology, Pakistan
[*] Department of Mechatronics Engineering, International Islamic University Malaysia
hasan.mohdzaki@research.uwa.edu.au, faisal.shafait@seeks.edu.pk, ajmal.mian@uwa.edu.au

Abstract

First-person videos have unique characteristics such as heavy egocentric motion, strong preceding events, salient transitional activities and post-event impacts. Action recognition methods designed for third person videos may not optimally represent actions captured by first-person videos. We propose a method to represent the high level dynamics of sub-events in first-person videos by dynamically pooling features of sub-intervals of time series using a temporal feature pooling function. The sub-event dynamics are then temporally aligned to make a new series. To keep track of how the sub-event dynamics evolve over time, we recursively employ the Fast Fourier Transform on a pyramidal temporal structure. The Fourier coefficients of the segment define the overall video representation. We perform experiments on two existing benchmark first-person video datasets which have been captured in a controlled environment. Addressing this gap, we introduce a new dataset collected from YouTube which has a larger number of classes and a greater diversity of capture conditions thereby more closely depicting real-world challenges in first-person video analysis. We compare our method to state-of-the-art first person and generic video recognition algorithms. Our method consistently outperforms the nearest competitors by 10.3%, 3.3% and 11.7% respectively on the three datasets.

1. Introduction

Video based human action recognition has received considerable attention from the research community due to its importance in real-world applications such as surveillance and human-machine interaction. The key challenge of designing highly discriminative motion features that can capture high-level video dynamics still remains an open problem. While a number of research works have focused on action recognition from a third-person's perspective, efforts on understanding first-person videos are much more restricted despite that wearable cameras are increasingly accessible. Moreover, many real applications can benefit from the understanding of such videos including medical monitoring, life-logging and long-term robotics perception [13, 16, 20, 30, 32, 33, 48, 45, 49, 51].

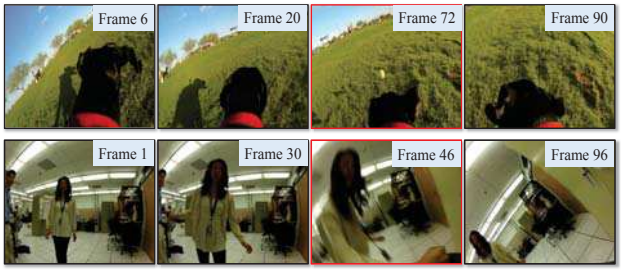

Figure 1. Sample frames of *play ball* and *punching* actions from our proposed YoutubeDog (top) and JPL-Interaction [32] (bottom) datasets respectively. First-person videos are defined by multiple salient sub-events such as pre-event (waiting/ standing), transitional (ball is being thrown/ approaching), action is being performed (playing ball/ punching), post-event (ball is released/ observer is knocked down). We model the dynamics of these sub-events before temporally encoding these dynamics as the entire video representation.

In contrast to third-person videos, where the camera is usually set up in a fixed position, in first-person videos the camera is moving and the observer is actively involved in the activities being recorded. The activities include inward interactions (*e.g.* shaking hands or throwing something to the observer), egocentric actions (*e.g.* jumping) and outward interactions (*e.g.* following a leader). This results in unique visual characteristics of first-person videos (see Fig. 1 for examples) that are not normally found in third-person videos. The visual effects can occur due to heavy egocentric motion (*e.g.* the person is running), strong preceding events (*e.g.* sudden approach before harming the observer), salient transitional activities (*e.g.* transition between running and playing ball) and post-event impacts (*e.g.* observer knocked down after being punched by an attacker). Moreover, most third-person videos [28, 43] consist of a subject performing an action repeatedly where the

discriminative pattern can be effectively captured. On the other hand, the motion dynamics in first-person videos do not exhibit such patterns. Conventional methods for third-person video representation such as global descriptors [36], densely sampled local features [39] or frame-wise representation [5, 28] may not adequately capture these factors that characterize first-person videos.

Recently, methods that are based on frame-wise features have gained some success in modeling temporal information due to the discriminative properties of generic feature extractors *i.e*. convolutional neural networks (CNN) [5, 10, 11, 26, 28, 50, 52]. This especially holds for third-person videos (*e.g*. multi-view surveillance cameras [27]) where the subject performs actions with a continuous pattern and is captured from a relatively static position. On the contrary, many first-person videos are more appropriately described by a set of short-term highlights and there are strong causal relationships between these events that formulate the semantics of the overall action [25, 41, 53]. In this paper, we demonstrate that these types of videos can be better represented when sub-event dynamics of the entire sequence are considered separately and a hierarchy is extracted over temporal domain of these sub-events to model the evolution of such dynamics.

In the proposed method, we first map each frame in the sequence to a discriminative space by forward passing through a CNN model and taking the first-fully connected layer activations. We consider each activation neuron as a point that changes according to a function of time *i.e*. as a time series. Then, we sample the temporal domain of the video into multiple overlapping segments. To capture the dynamics of these segments, we employ a temporal feature pooling function on the neurons in each defined interval to represent how the appearance of the actions evolve in short-term intervals. These sub-event dynamics are then temporally aligned and a group of Fourier coefficients are extracted in a temporal pyramid to encode the overall video representation. We show that our proposed method consistently outperforms other video representation techniques for first-person video recognition on two benchmark datasets and a newly proposed dataset of first person-videos. The compared video representation techniques include those that are specifically tailored for action recognition such as improved Dense Trajectories (iDT) [39], frame-based representation [10] and state-of-the-art features for first-person video recognition [33].

2. Related Work

First person video recognition is a relatively new area. Existing first-person action recognition methods can be categorized into two types. The first category methods focus on the application perspective and address the issues related to first-person vision whereas methods in the second category address the core challenge in action recognition *i.e*. designing feature representations for action classification. Here we present an overview of both research streams.

First-Person Action Recognition: First-person action recognition answers one of the two questions based on the *subject* of the action: "what are they doing to me?" [13, 30, 32, 45] or "what am I doing?" [16, 20, 33]. From practical application point of view, answering both questions is essential for first-person video recognition. However, existing techniques address only one of the questions. Moreover, they benchmark the algorithm on videos obtained in a controlled environment with limited number of subjects, cameras and environments [13, 32]. Therefore, the effectiveness of the existing methods [13, 16, 20, 30, 32, 33, 45] cannot be established for a realistic environment that contains a wide variety of action classes, subjects, environments and camera settings. In this paper, we benchmark our algorithm against several state-of-the-art methods for answering both questions. We also propose a new dataset of uncontrolled first-person videos containing more action classes, subjects, environmental variations, and cameras of different qualities.

Most of the literature in first-person video recognition focuses on long-term life logging within a specific environment (*e.g*. kitchen) and a limited set of applications such as hand detection and segmentation, egocentric object recognition or contextual face recognition [25, 41, 53]. There is also an increasing amount of literature that addresses the tasks of retrieving highlights/ important snapshots from long duration first-person videos [14, 22, 24, 46]. This paper addresses the core task of labelling actions in video clips. Our method is based on a moving window and can be applied to the above-mentioned recognition problems as well as other related tasks which are meant for long duration videos such as action detection and prediction [29].

Action Representation in Videos Hand-crafted features such as those that are based on spatio-temporal interest points [21] were dominant in the earlier works on video classification [36, 39, 40]. Due to the remarkable success of CNN on image classification, several methods have been proposed where deep networks are trained for video-based action recognition [5, 8, 17, 19, 35]. However, these methods are not specifically designed for first-person videos. Fine tuning such methods for first person video classification requires large-scale annotated data which is currently not available. On the other hand, frame-based representation has shown promising performance for action recognition [5, 10, 11, 26, 28, 52]. Qi *et al*. [26] extracted the fully connected layer activations of CNN as single frame features and showed that by simply taking the first- and second-order statistics of the frame-wise features, dynamic texture and scene classification can be significantly improved. There is also evidence that learning end-to-end frame based network shows comparable performance to learning multi-

frame based models [8, 19]. Our method goes beyond using CNN activations as per-frame features by deploying various temporal pooling operators to capture the dynamics of the segments which are used as the description of the sub-events in first-person videos. Moreover, as our proposed algorithm models the temporal representation at multiple scales and granularity, it is more closely related to the methods that employ hierarchical temporal structure [6, 9].

3. Proposed Methodology

In this section, we present our approach for designing a motion representation that is able to encode high-level abstraction of temporal structure for first-person videos. Our approach consists of core modules including the modeling of sub-event dynamics and a global temporal encoding.

3.1. Sub-Event Dynamics

Harvesting discriminative high-level motion features which lie on a non-linear manifold pose a key challenge in action recognition. Recent research works have shown that temporal information can be exploited in order to capture such dynamics, usually by employing recurrent based networks [5, 47]. Instead of taking the dynamics of the entire video directly, our approach starts by extracting multiple overlapping segments using pre-defined temporal filters from the whole video. These segments represent sub-events that are contained in each action video. Our main motivation is to observe the changing behaviour of the appearance in the frames over time, therefore revealing the motion dynamics in the short interval. This is done by applying a temporal feature pooling function (where we detail the choice of function) to capture different temporal statistics and features.

Concretely, let us assume a set of n labeled training videos, $\mathcal{V} = \{V^1, V^2, \dots, V^n\}$, each is assigned with one of the labels from C classes. Each video instance V^i can be represented by a set of ordered continuous frames $V^i = \langle v_1^i, v_2^i, \dots, v_{m_i}^i \rangle$, where $v_j^i \in \mathbb{R}^d$. From these sequence of vectors, we densely extract the sub-segments using a defined window size w and stride s for $t^s \in \{1, s+1, 2s+1, \dots\}$ where t^s denotes the starting frame number of the local time interval, *i.e.* $\{[t_1^s, t_1^e], \dots, [t_N^s, t_N^e]\}$. Our proposal maps these resultant sub-segments $\mathcal{S} = \{S_\ell\}_{\ell=1}^N$ onto a discriminative space (*i.e.* by feed-forwarding through CNN) to convert from $v_j \rightarrow x_j \in \mathbb{R}^D$, where $D = 4096$ is the dimension of the first fully-connected layer neurons activation of the CNN.

In each time interval, the sequence of w vectors of $x_1, x_2, \dots, x_w$ can now be interpreted as points that are changing according to the function of time. Let $x_j = \{x_j^1, x_j^2, \dots, x_j^D\}$ be the frame-wise feature descriptor extracted at frame j and that each x_j^i is a neuron from the CNN activation. Thus, we can now leverage each neuron in the activation vector as a time series, each with the dimension of w *i.e.* $f^k(t) = \{f^1(t), f^2(t), \dots, f^D(t)\}$. For each time series in a sub-segment, a temporal feature pooling function Ψ is invoked and this will generate a value for each time series. As a result, for any type of temporal pooling function, we get a D-dimensional vector for each sub-segment which is temporally aligned with other pooled features from other segments before the invocation of global temporal encoder to get the final feature representation.

3.2. Temporal Feature Pooling Functions

Here, we discuss multiple types of temporal feature pooling functions Ψ that are used to aggregate the frame-based descriptors within the pre-defined segments. We investigate the following functions in this paper: max pooling, sum pooling, histogram of time series gradient, cumulative histogram of time series gradient [33] and evolution pooling [10].

Max pooling, Ψ^{max}: Given a time series $f^k(t)$, the max pooling operator find the peak value in the series which is defined as

$$\theta_k^{\max}[t^s, t^e] = \max_{t=t^s,\dots t^e} f^k(t) \tag{1}$$

Sum pooling, Ψ^{sum}: Similarly, sum pooling aggregates all entries in the time series which can be expressed as

$$\theta_k^{\sum}[t^s, t^e] = \sum_{t=t^s}^{t^e} f^k(t) \tag{2}$$

Histogram of time series gradient, $\Psi^{\delta_1^+}, \Psi^{\delta_1^-}$: As the time series of CNN neurons are already in a discriminative space and we observe that the trends in time series of an action resembles those with the same label, we encode the information of how the trends move by calculating the number of positive and negative gradients within the time intervals. This operators will produce two values for each time series interval.

$$\begin{aligned}\theta_k^{\delta_1^+}[t^s, t^e] &= |\{t | f^k(t) - f^k(t-1) > 0 \wedge t^s \leq t \leq t^e\}| \\ \theta_k^{\delta_1^-}[t^s, t^e] &= |\{t | f^k(t) - f^k(t-1) < 0 \wedge t^s \leq t \leq t^e\}|\end{aligned} \tag{3}$$

Cumulative histogram of time series gradient, $\Psi^{\delta_2^+}, \Psi^{\delta_2^-}$: We also investigate a variant of Eq. 3 by taking the total number of positive and negative gradients in each time interval.

$$\begin{aligned}\theta_k^{\delta_2^+}[t^s, t^e] &= \sum_{t=t^s}^{t^e} h_k^+(t) \\ \theta_k^{\delta_2^-}[t^s, t^e] &= \sum_{t=t^s}^{t^e} h_k^-(t)\end{aligned} \tag{4}$$

where

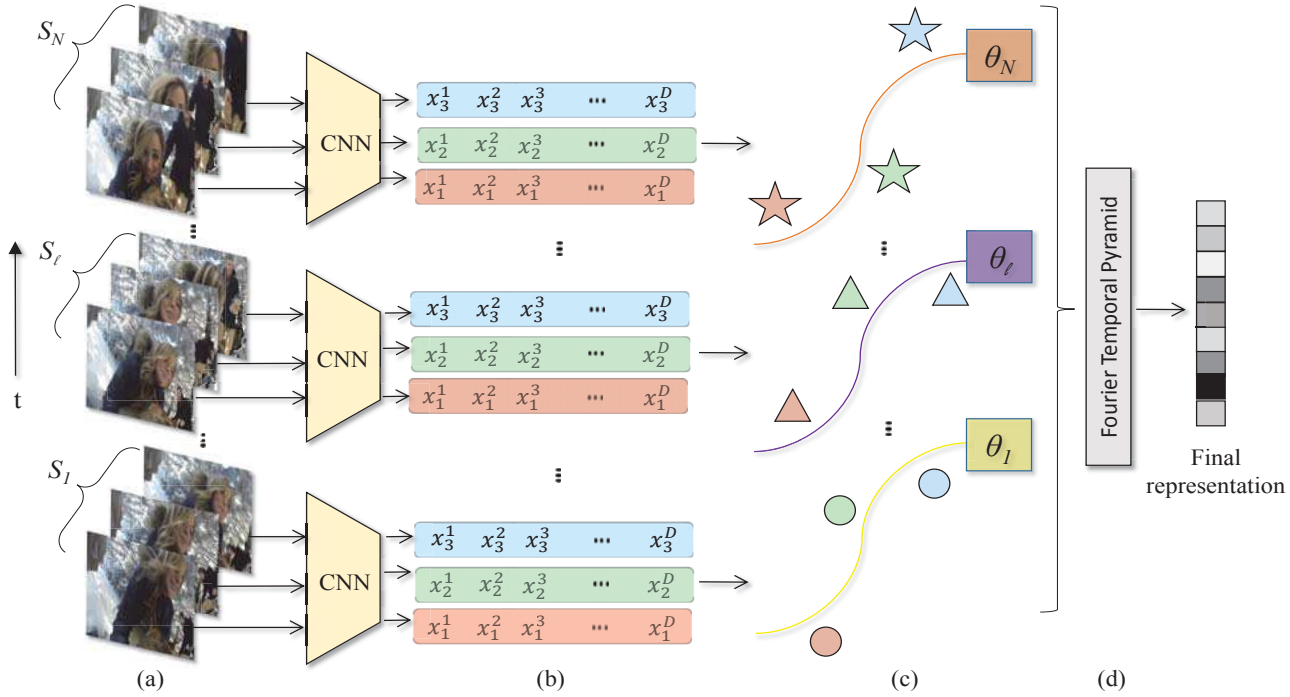

Figure 2. An overview of the proposed sub-event dynamics representation. (a) Video sequence is divided into overlapping sub-segments (in this example, with a window size of 3). (b) Each frame within sub-segments is individually passed through a CNN model and the fc_6 layer's activations are extracted. (c) A temporal feature pooling function is applied to the time series of each neuron activation (in this example, rank pooling [10] is illustrated) and pooled vectors are temporally aligned. (d) A Fourier Temporal Pyramid algorithm is applied to each pooled time series and concatenated to produce the final feature representation.

$$h_k^+(t) = \begin{cases} f^k(t) - f^k(t-1) & \text{if } (f^k(t) - f^k(t-1)) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$$h_k^-(t) = \begin{cases} f^k(t-1) - f^k(t) & \text{if } (f^k(t) - f^k(t-1)) < 0 \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Evolution or Rank Pooling Ψ^{rank}: So far we have discussed the temporal pooling functions which are *orderless*, that is they completely ignore the temporal order of the frames in the sequence. However, encoding temporal order information has shown to be an important feature for video representation as it can effectively model the evolution of the frame appearance [10, 11]. Therefore, we propose to encode the sub-segment dynamics by employing a rank pooling algorithm [10] on the CNN activations of the frames within the time intervals. Specifically, we are interested in getting a ranking machine to model a regression space which explicitly imposes a strict criterion that earlier frames must precede the following frames. To encode the dynamics of a sub-segment, the ranking function $\Psi^{rank}(f(t), \phi)$ parametrized by ϕ sorts the sequence frame such that $\forall t-1, t,\ f(t-1) \prec f(t) \iff \phi.f(t-1) < \phi.f(t)$. Using the principle of RankSVM [18] for structural risk minimization and max-margin formulation, the objective function can be expressed as

$$\begin{aligned} &\arg\min_{\phi} \frac{1}{2}\|\phi\|^2 + C \sum_{\forall a,b,\mathbf{v}_{ia} \prec \mathbf{v}_{ib}} \epsilon_{ab} \\ &s.t.\ \phi^T.(f(t) - f(t-1)) \geq 1 - \epsilon_{ab} \\ &\epsilon_{ab} \geq 0. \end{aligned} \tag{7}$$

To estimate this linear function and learn the parameter ϕ, we use the Support Vector Regressor (SVR) [23] for its efficiency. The learnt parameter ϕ defines the pooled feature representation for each sub-segment $\{\theta_\ell^{rank}\}_{\ell=1}^N$.

3.3. Global Temporal Encoding and Classification

After the sub-event dynamics have been encoded, we need to encode the global representation of the entire video. As the actions in first-person videos are performed with varying speeds and are captured in variable length clips, the intra-class actions exhibit highly misaligned temporal information which is a challenging nuisance for action matching. Hence, we propose to encode the global structure of the videos by employing Fourier Temporal Pyramid (FTP) [42] on the sub-segment dynamics. To do this, we first recursively divide the pooled sub-segment representations into a pyramid with $l = 1, \ldots, L$ levels, and perform

fast Fourier transform (FFT) on all the partitions to extract fine-to-coarse temporal structure.

Let $\theta^i_{k,\ell}$ represent each entry of the sub-segment dynamic features for the i-th action video, where $k = 1, \ldots, D$ denotes the index of the features entry and $\ell = 1, \ldots, N$ represents the frame number. At the lowest pyramid level, we perform FFT on $\theta^i_k = [\theta^i_{k,1}, \theta^i_{k,2}, \ldots, \theta^i_{k,N}]$ and get the first p low frequency coefficients. The same procedure is applied at each partition, and the concatenation of all low-frequency coefficients at all partitions is used as the FTP features for each entry. Finally, we define the spatio-temporal representation of the i-th video as the concatenation of the FTP features from all entries.

It is worth noting that our proposed video representation is completely unsupervised and is dictionary-free. This alleviates the need for re-training whenever data for the same problem from a new domain has to be processed. The only step that requires supervision is during the learning of an action classifier where we employ a linear Support Vector Machines (SVM) [7] to predict the action class. Hence, as long as action classes remain the same, re-training is not necessary.

4. Datasets

In this section, we provide an overview of the publicly available datasets for first person activity recognition and identify a common limitation of these datasets: controlled environmentalu setup. We overcome this gap by developing a new dataset that is collected from real-world videos and annotated according to action classes that occur in practical scenarios.

4.1. DogCentric Activity Dataset

This dataset [16] consists of first-person videos that were captured using wearable camera mounted on dogs. It has 10 actions from four subjects which capture the ego-centric behaviour of the dog as well as the interaction between the dog and human. The first-person actions include: (1) *waiting car*, (2) *drinking*, (3) *feeding*, (4) *looking left*, (5) *looking right*, (6) *patting*, (7) *playing ball*, (8) *shaking body*, (9) *sniffing* and (10) *walking*.

4.2. JPL First-Person Interaction Dataset

This dataset [32] consists of 57 continuous video sequences of actions done on the observer (*i.e*. a humanoid robot). There are 7 actions performed in 12 sets from 8 subjects in various indoor environment and lighting conditions. Sample snapshots of each action performed by different subjects are shown in Fig. 3.

4.3. YouTubeDog Action Dataset

Both of the above datasets have been collected in a controlled environment (*e.g*. in indoor laboratory, same subjects, same camera, same background). As first-person

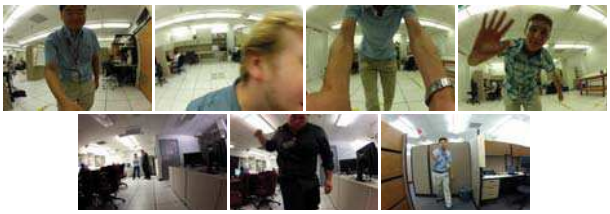

Figure 3. Sample snapshots from first-person videos in JPL-Interaction dataset [32]. From top (left) to bottom (right): *hand shake*, *hugging*, *petting*, *waving*, *point-converse*, *punching* and *throwing*.

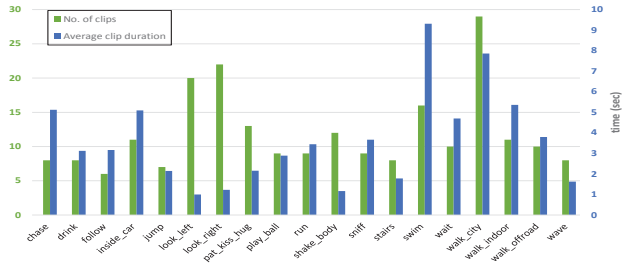

Figure 4. Total number of clips (green bars) and average clip duration (blue bars) for each class in our YouTubeDog dataset.

videos are typically used in an adverse outdoor environment, it is crucial to evaluate recognition algorithms in an uncontrolled environment (videos captured "in the wild") to establish their robustness. Therefore, we develop a new first-person dataset comprising video clips which were extracted from first-person videos in a public video sharing portal (*i.e*. YouTube). We make this dataset publicly available for research community [1].

Specifically, we collected videos captured by wearable cameras mounted on dogs. In contrast to the DogCentric dataset [16], there is no specific criterion used for the data capturing. Thus, the dataset contains videos from various backgrounds, dog species, camera quality and also the placement of the camera. We organized the videos into 226 short clips and annotated each clips into one of 19 action classes including: (1) *chase*, (2) *drink*, (3) *follow*, (4) *inside car*, (5) *jump*, (6) *look left*, (7) *look right*, (8) *pat/ kiss/ hug*, (9) *play ball*, (10) *run*, (11) *shake body*, (12) *sniff*, (13) *stairs*, (14) *swim* , (15) *wait traffic*, (16) *walk city*, (17) *walk indoor*, (18) *walk offroad* and (19) *wave*. This dataset is challenging as some videos are in a very low resolution and the appearance of the frames does not necessarily reflect the action being performed (*e.g*. frames show unimportant background while the dog is drinking water). Therefore, the algorithm has to exploit the temporal structure to describe the action. Sample snapshots from this dataset are shown in Figure 5. The chart shown in Fig. 4 illustrates the class-specific total number of clips and average clip duration.

[1] http://www.csse.uwa.edu.au/∼ajmal/databases.html

Figure 5. Sample snapshots from the first-person videos in our YouTubeDog dataset which has 19 action classes and diverse data distributions. From top (left) to bottom (right): *chase*, *drink*, *follow*, *inside car*, *jump*,*look left*, *look right*, *pet/kiss/hug*, *run*, *shake body*, *sniff*, *stairs*, *swim*, *wait traffic*, *walk city*, *walk indoor*, *walk offroad* and *wave*. The sample frames from the action *play ball* are depicted in Fig. 1.

5. Experiments

In this section, we evaluate our proposed approach on two distinctive types of recognition tasks; the actions performed by the observer and the actions done towards the observer. The former is represented by DogCentric Activity dataset [16] and YouTubeDog action dataset whereas the latter is encompassed by JPL-Interaction dataset [32]. The baseline results are reported from the original literature [16, 32, 33], the publicly available implementation or from our own careful re-implementation of the proposed methods.

Our main objective is to evaluate the proposed sub-event dynamics encoding technique for first-person action recognition. We use a CNN model which was discriminatively trained on ImageNet [3] dataset without any fine-tuning. Specifically, we feed-forward each video frame through the VGG-f [1] model and extract the activations of the first-fully connected layer fc_6. For all experiments, unless otherwise specified, we empirically set the number of FTP levels $L = 3$, and the number of low frequency Fourier coefficients $p = 4$ using cross-validation on training data. For rank pooling algorithm, we set the value of controlling parameter $C = 1$ as suggested by Fernando *et al*. [10]. Furthermore, except for the model ablation experiments (Sec. 5.1), we set the local window size w and stride size s to 15 and 2 respectively.

Hereinafter, we refer to our proposed sub-event dynamics modeling technique and global temporal encoding as SeDyn and FTP, respectively. For each dataset, we report the classification accuracy of our proposed SeDyn in two settings: (1) SeDyn, where we compute the mean of the sub-event dynamics (computed using Rank Pooling) and use the pooled vector as the feature representation of the video. (2) SeDyn+FTP, where we apply the proposed global temporal encoding on top of the SeDyn features to capture the entire video representation. Moreover, we re-implemented three baseline methods to examine the effectiveness of our proposed algorithm. These include: (1) Improved Dense Trajectory (iDT) [21] which is encoded into bag-of-word representation with 2000 visual words, (2) Pooled Time Series (PoT) with the same setting as Ryoo and Matthies [30] (temporal pyramid of 4 and four temporal operators) using the output of fc_6 of VGG-f, as well as (3) directly applying FTP on the per-frame fc_6 features without the sub-event dynamics. The reported accuracy of our proposed method for comparative analysis is taken from the best performing modules in Sec. 5.1.

Table 1. Comparing various temporal pooling functions for modeling sub-event dynamics in first-person action recognition datasets.

Method	DogCentric	JPL	YoutubeDog
Max Pooling	72.5	59.6	60.6
Sum Pooling	70.1	47.9	61.1
Gradient Pooling	67.9	83.0	56.3
Cumulative Gradient	72.7	90.2	64.4
Rank Pooling	**75.2**	**92.9**	**64.6**

5.1. Analysis of Individual Components

We examine different modules of the proposed sub-event dynamics representation in terms of their effect on the recognition accuracy. We start by experimenting with two hyper-parameters: (1) stride size s *i.e.* how many sub-events should be sampled and (2) local window size w *i.e.* the time interval of sub-events. For the evaluation of different stride sizes, we set a constant value of window size $w = 15$ while for the evaluation of different window sizes, we set the value of the stride $s = 2$. For both experiments, we apply Rank Pooling to capture the dynamics of sub-events and FTP to get the final video representation. Generally, the accuracy does not degrade if the amount of extracted sub-events is neither too large or too small. We used the training partition of the DogCentric dataset to tune these hyperparameters to a stride of 2 and a window size of 15 and used the same values for the remaining two datasets to show the generalization ability of our method.

5.1.1 Temporal Pooling

We conduct an experiment to show that the evolution/rank pooling method is the best for representing sub-event dynamics. Table 1 compares the rank pooling with other pooling functions for the first-person action recognition datasets. Max and sum pooling show similar performances as these pooling operators essentially capture the average behaviour of the neurons in the fc_6 features. Gradient and cumulative gradient pooling recorded a significant increase in accuracy

from that of max and sum pooling in JPL-Interaction dataset whereas the former method does not perform equally impressive for DogCentric and YoutubeDog datasets. This shows that while it is adequate to model the dynamics of sub-events in the actions done on the observer considering the quantized time series pattern (using gradient pooling), modeling the dynamics of sub-events in the actions done by the observer with various ego-motions needs to also consider the more precise location of the occurrences of the dynamics pattern. Nevertheless, employing Rank Pooling to model the sub-event dynamics gives the best performance for all datasets which suggests that such dynamics can be better captured when the evolution of the appearance in pre-defined time intervals is explicitly modeled. Moreover, while orderless pooling methods have proven successful for spatial based recognition such as image classification (*e.g.* [12]), it is more effective to pool multiple features in a more structured scheme by considering the ordered indices of the frames in the video sequences.

5.1.2 Combination with Trajectory Features

Recent action recognition research [8, 9, 38, 44] has suggested that deep network based features can be complemented with hand-crafted trajectory features (iDT) [39] to boost the performance. The improvement in accuracy comes from the fact that the learned features do not optimally capture the information in videos. To show that the proposed SeDyn is an optimal first-person video representation, we conduct a simple experiment by augmenting the SeDyn features with iDT [39] before classification. With the augmented features, the recognition accuracy does not improve on any dataset except the newly proposed dataset where the improvement is only 0.5%. This shows that our SeDyn is a more comprehensive representation.

5.2. Comparing Results on DogCentric Dataset

We follow the experimental protocol of Iwashita *et al*. [16]. Specifically, we randomly split the videos in each action class into training and testing dataset with 0.5 probability for 100 rounds. If the total number of videos is odd, we ensure that the testing set contains more number of videos. The final accuracy is taken as the average over the splits.

As depicted in Table 2, our SeDyn representation significantly outperforms all state-of-the-art and baseline methods. Local video feature based representations which are successful in recognizing third person videos [4, 21, 39] have proven insufficient for first-person videos. This is because first-person videos contain various dynamics and severe ego-motion that cannot be properly tracked by local features in consecutive frames. In addition, the performances of fc_6+FTP and PoT show that encoding the global structure of the entire video is important for first-person action recognition especially when we have rich frame-wise feature representation in hand such as CNN features.

Table 2. Performance comparison of the proposed techniques with state-of-the-art methods for first-person action recognition task on DogCentric dataset [16].

Method	Accuracy
STIP+IFV [21]	57.6
Cuboid+IFV [4]	59.6
Iwashita *et al*. [16]	60.5
PoT [33]	64.9
iDT [39]	52.9
PoT [33]+VGG-f	57.8
Proposed methods	
fc_6+FTP (baseline)	68.1
SeDyn	73.6
SeDyn + FTP	**75.2**

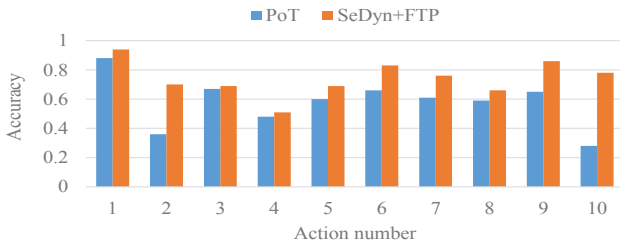

Figure 6. Class specific action recognition accuracies of PoT [33] and our proposed SeDyn+FTP on DogCentric Dataset [16].

However, these frame-wise based representations ignore the salient sub-event features that define the entire first-person actions which is important especially for actions with minimal appearance cues, as shown in Fig. 6. In contrast, our proposed SeDyn+FTP recorded highest accuracy of 75.2% as our algorithm is able to model a set of chronology between the salient sub-events and use this information to encode the global temporal video structure.

5.3. Comparing Results on JPL Dataset

The experimental procedure of Ryoo and Matthies [30] is adopted for this dataset. In particular, half action sets are used as the training data while testing is done on the remaining. This training/ testing split is sub-sampled 100 times and the mean accuracy is taken as the final result. We tabulate the results and per-class accuracy in Table 3 and Fig. 7 respectively.

Similarly to the previous dataset, local feature based representations are not able to effectively discriminate the action labels. However, there is a different performance pattern of the settings of our proposed SeDyn compared to those recorded in DogCentric dataset. Taking the simple average of sub-event dynamics does not give as good a performance as taking the global temporal encoding from the frame-wise features. This shows that sub-event moments in the actions done on the observer are generally sparse and that using the SeDyn alone without modelling the global

Table 3. Performance comparison of the proposed techniques with state-of-the-art methods for first-person action recognition task on JPL-Interaction data [32].

Method	Accuracy
ST-Pyramid [2]	86.0
Dynamic BoW [31]	87.1
Structure Match [32]	89.6
iDT [39]	60.5
PoT [33]+VGG-f	62.8
Proposed methods	
fc_6+FTP (baseline)	78.6
SeDyn	74.4
SeDyn+FTP	**92.9**

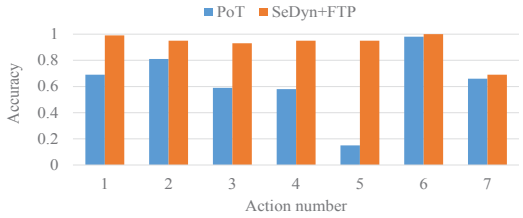

Figure 7. Class specific action recognition accuracies of PoT[33] and our proposed SeDyn+FTP on JPL Dataset [16].

temporal structure cannot adequately describe the action pattern. It is particularly true for actions such as *hand shake* and *throwing* where the ego-motion generally peaks only at specific time lapses.

Nevertheless, our proposed SeDyn+FTP considerably improves the state-of-the-art that shows first-person actions can be better modeled using both the sub-event dynamics and the global temporal structure encoding of these dynamics. It is important to emphasize that we do not fine-tune the CNN model to the target datasets which has shown improved performance for various applications [34]. Thus, state-of-the-art off-the-shelf per-frame features that possess effective discriminative properties [15, 37] can be employed in our method to further improve performance.

5.4. Comparing Results on YoutubeDog Dataset

Table 4. Performance comparison of the proposed techniques with the state-of-the-art methods for first-person action recognition task on our newly collected YoutubeDog Action dataset.

Method	Accuracy
iDT [39]	44.4
PoT [33]+VGG-f	52.9
RankPooling [10]	24.7
Proposed methods	
fc_6+FTP (baseline)	57.3
SeDyn	58.1
SeDyn+FTP	**64.6**

The results and class-specific accuracy on YouTubeDog action dataset are reported in Table 4 and Fig. 8 respectively. Our proposed SeDyn+FTP significantly outperforms the baseline methods by more than 10%. PoT representation [30] which has shown good performance for first-person videos is significantly lagging from our SeDyn. This demonstrates that first-person videos with severe ego-motion can be more discriminatively modeled using densely extracted sub-events instead of conventional temporal pyramid based representation used in PoT.

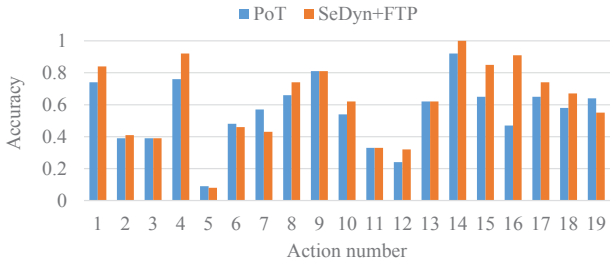

Figure 8. Class specific action recognition accuracies of PoT[33] and our proposed SeDyn+FTP on our YoutubeDog dataset.

Additionally, we also re-implemented Rank Pooling algorithm using the publicly available implementation [10, 11] that has shown impressive performance for action recognition in third-person videos. Interestingly, Rank Pooling on per-frame fc_6 features only registers 24.7% of accuracy, a massive degradation of 40% from our proposed SeDyn+FTP. This shows that modeling the appearance evolution of the video frames could not effectively encode discriminative temporal structure especially for sequences with short duration where the appearances may evolves very slowly and subtly.

6. Conclusion

We tackle the problem of action recognition from first-person videos by proposing a technique to model the sub-event dynamics. We build upon our observation that first-person videos are more appropriately described by the combination of salient sub-events. Using a set of overlapping sub-segments of an action video, we rank pool the segments to represent the sub-event dynamics and empirically show that such a pooling methodology consistently achieves the best performance over all datasets compared to conventional pooling methods. The global temporal structure of the video is then encoded with pyramidal Fourier coefficients. Experiments on two benchmark first-person video datasets as well as our newly proposed dataset show that our method consistently outperforms state-of-the-art by a significant margin.

Acknowledgements

The corresponding author is sponsored by Ministry of Education Malaysia and International Islamic University Malaysia. This research was supported by ARC Discovery Grant DP160101458. We also thank NVIDIA for donating the Tesla K-40 GPU.

References

[1] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *Proc. BMVC*, 2014. 6

[2] J. Choi, W. J. Jeon, and S.-C. Lee. Spatio-temporal pyramid matching for sports videos. In *Proc. of ACM international conference on Multimedia information retrieval*, 2008. 8

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. CVPR*, 2009. 6

[4] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*. IEEE, 2005. 7

[5] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proc. CVPR*, 2015. 2, 3

[6] Y. Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proc. CVPR*, 2015. 3

[7] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *JMLR*, 9(Aug):1871–1874, 2008. 5

[8] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proc. CVPR*, 2016. 2, 3, 7

[9] B. Fernando, P. Anderson, M. Hutter, and S. Gould. Discriminative hierarchical rank pooling for activity recognition. In *Proc. CVPR*, 2016. 3, 7

[10] B. Fernando, E. Gavves, J. Oramas, A. Ghodrati, and T. Tuytelaars. Rank pooling for action recognition. *PAMI*, 2016. 2, 3, 4, 6, 8

[11] B. Fernando, E. Gavves, J. M. Oramas, A. Ghodrati, and T. Tuytelaars. Modeling video evolution for action recognition. In *Proc. CVPR*, 2015. 2, 4, 8

[12] Y. Gong, L. Wang, R. Guo, and S. Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. In *Proc. ECCV*, 2014. 7

[13] I. Gori, J. Aggarwal, L. Matthies, and M. Ryoo. Multitype activity recognition in robot-centric scenarios. *IEEE Robotics and Automation Letters*, 1(1):593–600, 2016. 1, 2

[14] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool. Creating summaries from user videos. In *Proc. ECCV*, 2014. 2

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. CVPR*, 2016. 8

[16] Y. Iwashita, A. Takamine, R. Kurazume, and M. Ryoo. First-person animal activity recognition from egocentric videos. In *Proc. ICPR*, 2014. 1, 2, 5, 6, 7, 8

[17] S. Ji, W. Xu, M. Yang, and K. Yu. 3D convolutional neural networks for human action recognition. *PAMI*, 35(1):221–231, 2013. 2

[18] T. Joachims. Training linear SVMs in linear time. In *Proc. of ACM international conference on Knowledge discovery and data mining*, 2006. 4

[19] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proc. CVPR*, 2014. 2, 3

[20] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *Proc. CVPR*, 2011. 1, 2

[21] I. Laptev. On space-time interest points. *IJCV*, 2005. 2, 6, 7

[22] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *Proc. CVPR*, 2012. 2

[23] T.-Y. Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009. 4

[24] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *Proc. CVPR*, 2013. 2

[25] K. K. Minghuang Ma. Going deeper into first-person activity recognition. In *Proc. CVPR*, 2016. 2

[26] X. Qi, C.-G. Li, G. Zhao, X. Hong, and M. Pietikäinen. Dynamic texture and scene classification by transferring deep image features. *Neurocomputing*, 171:1230–1241, 2016. 2

[27] H. Rahmani and A. Mian. Learning a non-linear knowledge transfer model for cross-view action recognition. In *Proc. CVPR*, 2015. 2

[28] H. Rahmani and A. Mian. 3D action recognition from novel viewpoints. In *Proc. CVPR*, 2016. 1, 2

[29] M. Ryoo, T. J. Fuchs, L. Xia, J. K. Aggarwal, and L. Matthies. Robot-centric activity prediction from first-person videos: What will they do to me? In *Proc. on Human-Robot Interaction*, 2015. 2

[30] M. Ryoo and L. Matthies. First-person activity recognition: Feature, temporal structure, and prediction. *IJCV*, pages 1–22, 2016. 1, 2, 6, 7, 8

[31] M. S. Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *Proc. ICCV*, 2011. 8

[32] M. S. Ryoo and L. Matthies. First-person activity recognition: What are they doing to me? In *Proc. CVPR*, 2013. 1, 2, 5, 6, 8

[33] M. S. Ryoo, B. Rothrock, and L. Matthies. Pooled motion features for first-person videos. In *Proc. CVPR*, 2015. 1, 2, 3, 6, 7, 8

[34] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. In *Proc. CVPRW*, 2014. 8

[35] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Proc. NIPS*, 2014. 2

[36] B. Solmaz, S. M. Assari, and M. Shah. Classifying web videos using a global video descriptor. *Machine vision and applications*, 24(7):1473–1485, 2013. 2

[37] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proc. CVPR*, 2015. 8

[38] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3D convolutional networks. In *Proc. ICCV*, 2015. 7

[39] H. Wang and C. Schmid. Action recognition with improved trajectories. In *Proc. ICCV*, 2013. 2, 7, 8

[40] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *Proc. BMVC*, 2009. 2

[41] J. Wang, Y. Cheng, and R. Schmidt Feris. Walk and learn: Facial attribute representation learning from egocentric video and contextual data. In *Proc. CVPR*, 2016. 2

[42] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Learning actionlet ensemble for 3D human action recognition. *PAMI*, 36(5):914–927, 2014. 4

[43] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu. Cross-view action modeling, learning and recognition. In *Proc. CVPR*, 2014. 1

[44] L. Wang, Y. Qiao, and X. Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *Proc. CVPR*, 2015. 7

[45] L. Xia, I. Gori, J. K. Aggarwal, and M. S. Ryoo. Robot-centric activity recognition from first-person RGB-D videos. In *Proc. WACV*, 2015. 1, 2

[46] T. Yao, T. Mei, and Y. Rui. Highlight detection with pairwise deep ranking for first-person video summarization. In *Proc. CVPR*, 2016. 2

[47] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In *Proc. CVPR*, 2015. 3

[48] H. F. M. Zaki, F. Shafait, and A. Mian. Localized deep extreme learning machines for efficient RGB-D object recognition. In *Proc. DICTA*, 2015. 1

[49] H. F. M. Zaki, F. Shafait, and A. Mian. Convolutional hypercube pyramid for accurate RGB-D object category and instance recognition. In *Proc. ICRA*, 2016. 1

[50] H. F. M. Zaki, F. Shafait, and A. Mian. Modeling 2D appearance evolution for 3D object categorization. In *Proc. DICTA*, 2016. 2

[51] H. F. M. Zaki, F. Shafait, and A. Mian. Learning a deeply supervised multi-modal RGB-D embedding for semantic scene and object category recognition. *Robotics and Autonomous Systems*, 2017. 1

[52] S. Zha, F. Luisier, W. Andrews, N. Srivastava, and R. Salakhutdinov. Exploiting image-trained CNN architectures for unconstrained video classification. In *Proc. BMVC*, 2015. 2

[53] Y. Zhou, B. Ni, R. Hong, X. Yang, and Q. Tian. Cascaded interactional targeting network for egocentric video analysis. In *Proc. CVPR*, 2016. 2