Part-Activated Deep Reinforcement Learning for Action Prediction

Lei Chen[1], Jiwen Lu[2,*], Zhanjie Song[1], and Jie Zhou[2]

[1] Tianjin University, Tianjin, China
[2] Tsinghua University, Beijing, China
{chen_lei,zhanjiesong}@tju.edu.cn
{lujiwen,jzhou}@tsinghua.edu.cn

Abstract. In this paper, we propose a part-activated deep reinforcement learning (PA-DRL) method for action prediction. Most existing methods for action prediction utilize the evolution of whole frames to model actions, which cannot avoid the noise of the current action, especially in the early prediction. Moreover, the loss of structural information of human body diminishes the capacity of features to describe actions. To address this, we design the PA-DRL to exploit the structure of the human body by extracting skeleton proposals under a deep reinforcement learning framework. Specifically, we extract features from different parts of the human body individually and activate the action-related parts in features to enhance the representation. Our method not only exploits the structure information of the human body, but also considers the saliency part for expressing actions. We evaluate our method on three popular action prediction datasets: UT-Interaction, BIT-Interaction and UCF101. Our experimental results demonstrate that our method achieves the performance with state-of-the-arts.

Keywords: Action prediction, deep reinforcement learning, skeleton, part model

1 Introduction

Human activity analysis has aroused much attention in computer vision due to its broad prospects of applications [8, 17, 19, 29, 36]. As an important branch of human activity analysis, predicting the activity of humans presents importance in a number of real-world applications, such as video detection [43], abnormal behavior detection [7, 40] and robot interaction [36].In spite of the enormous amount of works conducted in this area [21, 25, 26], this task is still challenging due to the fundamental challenges inherent in the problem like the large variance among the activities and huge spatiotemporal scale variation. Recognizing the action in the full length video is too luxury to wait for the whole video completed [44]. For example, predicting the falling down of human can save the person as early as possible. Different from action recognition [4, 5, 34, 35, 41, 42, 46, 47],

* Corresponding author.

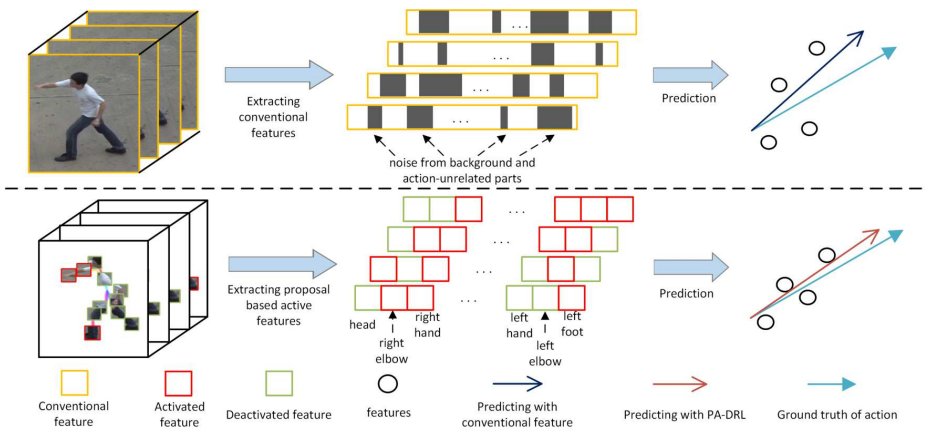


Fig. 1: **The advantage of PA-DRL in representing the evolution of the action.** Top: The features are extracted by the conventional methods with noise, which disturbs the action prediction. The gray parts in the feature are noise which are introduced by feature extraction and push the features far away from the evolution direction of the actual actions. Bottom: Our method is to purify the feature from noise with considering the structure of the human body. Moreover, we activate the action-related parts for action prediction to pull the feature of frames to the direction of action evolution. (Best viewed in color)

action prediction aims to take full advantage of the partially observed video for predicting the action as early as possible. Action prediction [31] is defined as the inference of the ongoing activities of humans just with observing the partial videos or sequences, when the action has not been completed.

It is challenging to model the process of the partially observed action in both spatial and temporal domain for predicting the ongoing actions [16]. The majority of the existing action prediction works can be mainly divided into two categories: exploiting reliable features with template matching [2,20,44] and developing classification models [14,16,31,33]. Approaches in the first category aim to design a template-based model for the prediction. However, these templates are easily affected by the outliers and perform poorly when actors present large pose variations. Methods in the second category focus on discovering temporal characteristics of human actions, because the confidence of prediction increases with more frames gradually observed. However, most existing methods extract holistic features of frames to exploit the temporal information, which ignores the essential structural information of the human body. The challenge of action prediction is that the useful information of predicting actions is very limited, but the redundant information has strong ability of disturbing the prediction. The top of Fig. 1 shows that the conventional features extracting from the whole frame captures the noise, which disturbs the action prediction.

To address the above limitations, we present a part-activated deep reinforcement learning method for action prediction. The bottom of Fig. 1 shows the main

process of activating the action-related parts. Based on the structural information of the human body by skeleton proposals, we activate the action-related parts of human body and deactivate the noise parts by deep reinforcement learning. Depending on the skeleton of the human body, our method extracts features in the region proposals which are decided by the joints of skeleton. Then we concatenate the features in order to keep the structural information of the body. For different actions, our method attends the action-related parts of features. Our proposed method learns a part-activated policy for activating and deactivating the parts of features with the deep reinforcement learning. Experimental results on three benchmarks demonstrate the effectiveness of our proposed approach.

2 Related Work

Action Prediction: Simply modeling action prediction as an ensemble of action classification is non-optimal. Conventional action recognition methods hold the assumption that the temporal information of an activity is complete, while only partial temporal information of an action is observed in action prediction. Most existing methods for action prediction can be divided into two categories: exploiting reliable features and developing classification models. For the first category, most existing methods design a template for action prediction. For example, Ryoo [31] proposed the integral bag-of-words (IBoW) and dynamic bag-of-words (DBoW) approaches for action prediction. The action model of every progress level is computed by averaging features of a particular progress level with the same category. The model suffers from the difficulty with the situation that the videos of the same action have a large variation in the spatial domain and it is sensitive to the outliers. Lan *et al.* [20] exploited templates at multiple levels of granularities in a hierarchical representation, which can capture and compare human movements at different context levels. For the second category, methods focus on exploiting the temporal information of human actions. For example, Cao *et al.* [2] designed an action prediction model with sparse coding to learn the features and reconstructed the testing partial videos with the bases extracted from the training activities. In their model, they addressed the problem of the intra-class action variations with bases from long-short segments. Kong *et al.* [16] proposed a multiple temporal scale support vector machine (MTSSVM) for action prediction and they took full advantage of the evolution of segments. Mahmud *et al.* [24] proposed a hybrid Siamese network with three branches to jointly learn both the future label and the starting time. They found that using more frames yielded high prediction performance. However, most existing methods try to capture the temporal information through the duration of partially observed actions, which ignore the importance of structural information of the human body for action representation.

Deep Reinforcement Learning: Recently, the field of reinforcement learning resurrects with the strong support from deep learning [13, 18, 45]. Deep reinforcement learning effectively learns the better policy than the supervised way for challenge tasks [22] and it can be divided into two main architectures:

Q-network and policy gradient. Deep reinforcement learning technique is introduced to optimize the sequential model with delayed reward [23] and performs very promising results in a series of problems. For example, Mnih *et al.* [28] achieved the human-level performance in the Atari game with their proposed deep Q-networks. Goodrich *et al.* [6] designed an architecture with 32 actions to shift the focal point and reward the agent when finding the goal. Caicedo *et al.* [1] defined a transformations set for the bounding box as the action of agent and rewarded the agent when the bounding box moves close to the ground-truth with iterations. More recently, deep reinforcement learning has been applied in many computer vision tasks [9–11,27,38]. For example, Krull *et al.* [18] applied a policy gradient approach to the object pose estimation problem. Kong *et al.* [13] proposed a novel multi-agent Q-learning solution that facilitates learnable inter-agent communication with gated cross connections between the Q-networks. Ren *et al.* [30] presented a novel decision-making framework for image captioning utilizing a policy network and a value network. However, little progress has been made in reinforcement learning for activity analysis, especially in action prediction. In this work, we develop a part-activated deep reinforcement learning model to learn the policy of activating the attention parts of the human body for predicting the unfinished actions.

3 Approach

In this section, we first show the pipeline of our part-activated deep reinforcement learning (PA-DRL) method for action prediction. Then we describe our proposed skeletal proposal for extracting features of actions. Lastly, we detail the method of our part-activated deep reinforcement learning method.

Fig. 2 shows the pipeline of our action prediction architecture and we first utilize the skeleton extracted by [3] to extract proposal. To take advantage of the information of the partial sequence, we provide the proposal for extracting the action features. To predict the action effectively, we design a two-step architecture to activate the original features extracted from the frames of videos.

- We extract those features in local patches of skeleton proposals which are decided by the joints of the skeleton. The extracted features contain action-related information and are used as the candidates for part-activating. Then we concatenate the features from the same frame in the order of the skeleton to keep the structural information of human body.
- We active the most related parts in features by learning the part-activated strategy with deep reinforcement learning. The activated parts of consequent frames enhance the action-related information in both spatial and temporal domain, which reduces the distance between the predicting action and the actual action in feature space.

3.1 Partial Feature Extraction

To solve the problem for the lack of apparent information of the skeleton, we use the skeleton as a proposal to select a local patch around the joint points.

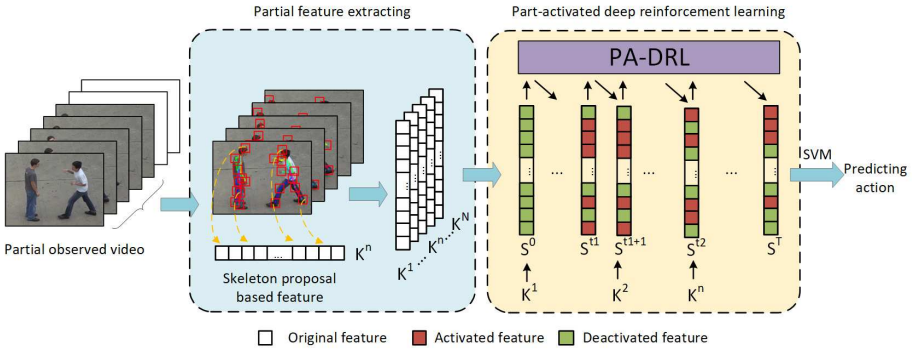


Fig. 2: **The pipeline of our PA-DRL.** The aim of our PA-DRL is to predict the ongoing actions based on the frames at the beginning of the action. The input are partially observed videos and sequences. We extract the skeleton proposal based features from every frame. Then our PA-DRL activates action-related parts of features frame by frame with deep reinforcement learning. The activated parts of features capture the action-related information, which are used for predicting the actions.

We extract apparent features in the local patches to provide spatial information for the structure of human body. The local patch extracted by joints of skeleton in our architecture is denoted as the skeleton proposal. The skeleton proposal carries two parts of information:

- The patches are extracted from images and contain the apparent information around the joints.
- The concatenating order of features keeps the structural information of human body.

All patches in the skeleton proposal are decided by the skeleton joints and the order of the concatenated feature keeps the structure of skeleton.

We define the operation of concatenating the features as $\Gamma(\cdot)$. We propose our skeleton proposal for video-based action prediction. Thus the input of our method are videos and sequences. For a video, the number of observed frames is defined as N. In every frame, the index of persons are indicated as p and the total number of people is P. The skeletons of people are defined by their indexes $\{S_1, S_2, ..., S_p, ..., S_P\}$. We assume that the skeleton has E joints that can be represented as $\{J_{S_p,1}, J_{S_p,2}, ..., J_{S_p,e}, ..., J_{S_p,E}\}$. Then we extract features based on joints of the skeleton, which is denoted by $F^n_{J_{p,e}}$. To keep the structure of skeleton, we concatenate the features in the order from 1 to E for one person. $U = P \times E$ represents the total number of parts in the state. For the nth frame, we formulate the concatenation of features as follows:

$$K^n = \Gamma_{p\in P}(\Gamma_{e\in E}(F^n_{J_{p,e}})) = \Gamma_{u\in U}(F^n_{J_u}), \tag{1}$$

where $F^n_{J_{p,e}}$ is the proposal based feature which is generated from the skeleton proposal. Comparing with the original features, the proposal-based feature not

only has less noise from background, but also keeps the structural information of actions. K^n is the concatenation of features denoting the representation for all persons in frame n, which has two advantages on representing the video:

- In K^n, the order of parts from $F^n_{J_{1,1}}$ to $F^n_{J_{P,E}}$ keeps the structural information of human body.
- The corresponding parts from K^1 to K^n captures the evolution of the corresponding human body parts.

Different parts in the same feature have different relationship to the action. The same part in different stages of the action has different significance in the whole action. For example, during the action of *boxing*, most of the joints of skeleton are moving. The joints at elbows and hands are positive for understanding the action of *boxing*. But the joints on feet disturb the representation of features. To address the problem of the noise from the unrelated parts for actions, we propose a part-activated deep reinforcement learning method to select the saliency parts of features on the human body.

3.2 Part-Activated Deep Reinforcement Learning

For action prediction, we use the partially observed videos to recognize the action. The number of observed frames is much less than the whole video. Our aim is to predict the ongoing action as early as possible. With a few frames at the beginning of the action, it is very essential to take fully advantage of action-related parts and to reduce interference of noise. We propose part-activated deep reinforcement learning method (PA-DRL) to active the action-related parts. The architecture of our method is based on the actor-critic. For every frame, there exists not "ground-truth" of action-related part. However, our PA-DRL is to learn the policy of activating the action-related parts only with the label of the action. Based on the deep reinforcement learning, PA-DRL makes a series of decisions to get the holistic optimal result for activating the action-related parts.

Fig. 3 shows the part-activated process of our method with observed frames. As shown in this figure, the red points represent the activated parts in the feature. With the starting frames of the action, the noise pushes the features away from the actual action in feature space, which makes the predicting evolutionary direction (yellow arrow) away from the actual evolutionary direction (black line of dashes). PA-DRL deactivates the parts with lager distance and pull features close to the actual action. The green points are the action-unrelated parts in features and are deactivated by our PA-DRL. The part-activated features predict a new evolutionary direction to represent the action, which is close to the actual action. PA-DRL deactivates some parts of the feature and changes the prediction result with deep reinforcement learning.

Problem Settings: To activate and deactivate parts in the feature, we have to confirm the relationship between parts and the action. However, it is hard to obtain all labels of activation and deactivation for every part in skeleton proposal based features. Different from conventional supervised deep learning methods,

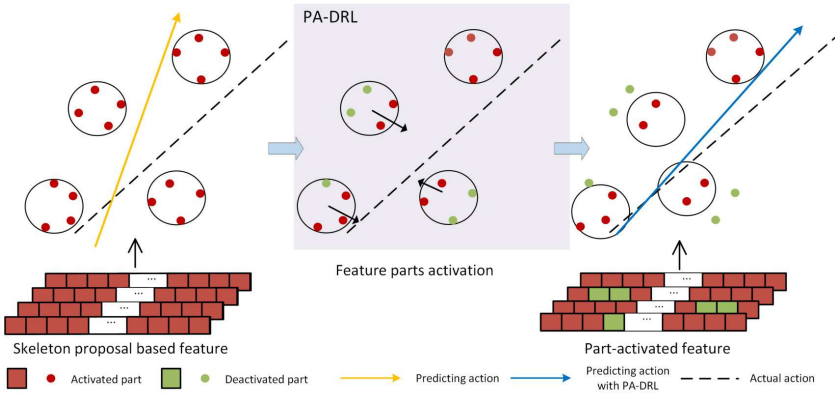


Fig. 3: **The PA-DRL for feature parts activation.** The red points represent the activated parts in corresponding features. The large black circle represents the whole feature which is decided by the all activated parts in the feature. The black circles from left to right reflect the temporal evolution of the action. Yellow arrow is the predicting evolutionary direction of the action. Black dashed line is the actual evolutionary direction of the action.

our PA-DRL aims to learn the policy for activating the action-related parts and deactivate the noise without labels of all parts. Based on the deep reinforcement leaning, our PA-DRL has three important elements: state, action and reward. To distinguish the action in the prediction task and the action in the learning architecture, we use *action* for the action in the learning architecture instead.

We define the *action* space Λ with two types of action for every part of state S_w^t. We denote the *action* $a_{u,w}^t \in \Lambda$ for the part $\beta_{u,w}^t$. Two types of *action* in action space Λ are activation and deactivation. For $\beta_{u,w}^t$, the *action* of activation can be represented by a vector of 1 with the same dimension of $\beta_{u,w}^t$. Similarly, the *action* deactivation can be represented by a vector of 0. Then we represent concatenated *action* A_w^t for state S_w^t as follows:

$$A_w^t = [a_{1,w}^t, ..., a_{U,w}^t], a_{u,w}^t \in \{1^b, 0^b\}, \tag{2}$$

where b is the dimension of feature $\beta_{u,w}^t$.

We define the state of our policy as S_w^t, where $w \in W$ is the index of videos and $t \in T_w$ is the iteration of learning process. T_w is the terminal step of video w. The original state S_w^0 equals the skeleton proposal based feature K^0. During the learning process, the state S_w changes with the iteration t. S_w^t denotes the activated feature at the tth iteration for the wth video. Thus we formulate the state S_w^t as follows:

$$S_w^t = \Gamma_{u \in U} \left(\beta_{u,w}^t\right), \tag{3}$$

where $\beta_{u,w}^t$ is the uth part of state S_w after the tth iteration.

We define the step reward for *action* A_w^t as $r(A_w^t)$. The predicting label of state S_w^t is η_w^t. The label of ground truth for the corresponding frame is ϵ_w. If η_w^t

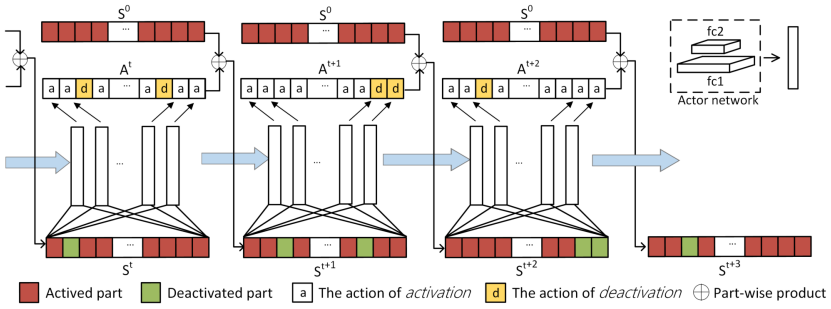


Fig. 4: **State transformation in PA-DRL.** The activated part is the corresponding part of original state. The deactivated part is the original part multiplying by 0. The deactivated part can be reactivated by adding the corresponding original part. We define a part-wise product to generate the new state. The new state S^{t+1} is computed by the part-wise product of the last state S^t and corresponding *action* A^t.

equals to ϵ_w, the *action* A_w^{t-1} is positive for prediction and will receive a positive reward by the policy. If the category of action is correctly predicted in continuous iterations, we decide a reward of $|r(A_w^{t-1})|+1$ for the *action*. Otherwise, we give the negative reward for the continuous wrong predictions with $-|r(A_w^{t-1})|-1$. At iteration t, we formulate the step reward $r(A_w^t)$ as follows:

$$r(A_w^t) = \Theta(\eta_w^t, \epsilon_w) \times (|r(A_w^{t-1})| + 1), \tag{4}$$

where $\Theta(\cdot)$ is characteristic function which equals 1 if prediction is correct and equals 0 else. Based on the step reward, we define the final reward function as $R(w)$. After the terminal iteration, we feedback the series *actions* A_w with the final reward $R(w)$. When the state stops at the terminal iteration S_w^T, the final reward $R(w)$ is the average value of all step rewards of every iteration. We represent final reward $R(w)$ for the video w as:

$$R(w) = \frac{1}{T}\sum_{t\in T} r(A_w^t). \tag{5}$$

The final reward is used for updating the model at the terminal step of one training sample.

State Transformation: Fig. 4 shows the state transformation of PA-DRL. We denote the actor network as Π_θ, which is parameterized by θ. We formulate the state transformation from state S_w^{t-1} to S_w^t as follows:

$$A_w^{t-1} = \Pi_\theta(S_w^{t-1}), \tag{6}$$

$$S_w^t = S_w^0 \odot A_w^{t-1}, \tag{7}$$

where S_w^0 is the state of original skeleton proposal feature and $\odot$ is element-wise product. The *action* A_w^{t-1} implements on the original state S_w^0 to activate and deactivate the parts of features. We prefer to compute the element wise product

of A_w^{t-1} and S_w^0. The reason is that the *action* A_w^{t-1} cannot reactivate the part of S_w^{t-1} which is already deactivated in the previous iterations. The information of the part is lost and cannot be recovered in the following iterations. After several iterations, the state of S_w^T has the possibility to form a vector of all zeros, which cannot represent the action.

To stop the iteration softly, we set the condition of terminal iteration. For the training process, we count the number of continuously correct predictions as σ. When σ is larger than the stop value λ, the iteration stops. For the testing process, we count the number of continuously consistent predictions as σ. The condition of termination is the same as that in training. The last *action* of frame n is the initial *action* of frame $n+1$. Because the adjacent two frames are similar, which makes the *action* similar. We use the constraint of continuity on the *action* to reflect the continuity of frames.

3.3 Implement Details

We first utilized the skeleton which was extracted with [3] as the proposal for extracting the features of the human body. The number of parts extracted by [3] is 14 and we uniformly used 28 parts (14×2) per frame, where we trained one actor network for each part. For every frame, we selected region proposals at joints of skeleton. The center of region was the joint of skeleton and the size of region was 20×20 pixels. At every joint, we extracted of the spatial feature with pre-trained model of VGG-16. The feature size of one patch was 1000. The extracted 14 features of one person were concatenated into a vector and then the features of different persons were concatenated. The generated feature was used as the input of our actor network. Our actor network consisted of two fully convolution layer $fc1$ and $fc2$. The $fc1$ layer had 128 units and the $fc2$ layer had 2 units. The layer of $fc1$ and $fc2$ were activated by the $relu(\cdot)$ function. We separately trained 28 actor networks with the same input and the output of every network referred to the *action* corresponding to the part. The critic part of our reinforcement learning was linear-SVM. The stop value λ was set as 5. The max value in training process and testing process were both set as 10. We used Adam [12] as the optimizer in training and set the learning rate as 10^{-4}. The discount factor γ was set as 0.99.

4 Experiments and Results

4.1 Datasets

We evaluated PA-DRL on the UT-Interaction #1, the UT-Interaction #2 [32], the BIT-Interaction dataset [15] and UCF101 dataset [37].

UTI #1 and UTI #2 datasets: The two sets of the UT-Interaction dataset contain videos of continuous actions of 6 classes: shake-hands, point, hug, push, kick and punch. Each video contains at least one execution per interaction, providing 8 executions of human activities per video on average. Both

sets have 60 video clips with 10 videos per action class. Backgrounds in the Set #2 are more complex than those in the Set #1.

BIT-Interaction dataset: The BIT-Interaction dataset has a list of 23 interactive phrases based on 17 attributes for all the videos. Videos are captured in realistic scenes with cluttered background. People in each interaction class behave totally different and thus have diverse motion attributes. This dataset consists of 8 classes of human interactions (bow, boxing, handshake, high-five, hug, kick, pat, and push), with 50 videos per class.

UCF101 dataset: UCF101 action recognition dataset has collected from YouTube, where the videos are realistic without constraint. The total categories of dataset are 101 and all videos are divided into 25 groups with 101 action categories. There are 13320 videos in all 101 categories.

4.2 Experimental Settings

In the training process, we trained the model by feeding one sample each round. We fixed the parameters of the linear-SVM in the process of training with each video. The model updated with the final reward by using (5) and calculated the reward with (4) in every iteration. In the training process, we terminated the iterative updating of each video when predicted labels were the same as the ground truth or the iteration number reached the max value. The parameters of network in our model updated after the iterative updating of each video was terminated. In the testing process, PA-DRL outputted the final feature for one video when predicted labels did not change in continuous 3 iterations or the iteration number reached the max value. On UT-Interaction datasets, we followed the experimental settings in [32] and utilized 10-fold leave-one-sequence-out to measure the performance of our proposal based PA-DRL on both the UTI #1 and the UTI #2. For every round, we measured the performance 10 times while changing the test set iteratively, finding the average performance. Every time, we utilized 6 videos in one of 10 folds as the testing set and used the other 54 videos as the training set. On BIT-Interaction dataset, we followed the settings in [15]. With randomly choosing 272 videos, we trained the model and utilized the remaining 128 videos for testing. On UCF101 dataset, we followed the split scheme proposed in [37]. We used the first 15 groups for training, the next 3 groups for cross-validation and the remaining 7 groups for testing.

4.3 Results and Analysis

We first compared our PA-DRL method with thirteen state-of-the-art action prediction methods, including SVM [31], Bayesian [31], IBOW [31], DBOW [31], SC [2], MSSC [2], Lan *et. al* [20], MTSSVM [16], AAC [44], MMAPM [14], C3D [39], Lai *et. al* [19] and Deep SCN [17]. We employed the results of these compared methods provided by the original authors. Table 1 illustrates the accuracy of PA-DRL compared with several state-of-the-art methods for action prediction. The comparisons were taken on UTI, BIT and UCF101 dataset at $OR = 0.5$ and $OR = 1.0$ separately. The OR indicates the observation ratio.

Table 1: The accuracy (%) of different methods on the UTI #1 , the UTI #2.

Methods	UTI Set #1		UTI Set #2	
	OR=0.5	OR=1.0	OR=0.5	OR=1.0
SVM [31]	25.3	69.2	27.2	69.2
Bayesian [31]	20.9	78.0	21.8	50.7
IBoW [31]	65.0	81.7	45.7	59.3
DBoW [31]	70.0	85.0	51.2	65.3
SC [2]	70.0	76.7	68.5	80.0
MSSC [2]	70.0	83.3	71.0	81.5
Lan *et. al* [20]	83.1	88.4	78.3	82.0
MTSSVM [16]	78.3	95.0	74.3	87.3
AAC [44]	88.3	95.0	75.6	63.9
MMAPM [14]	78.3	95.0	75.0	87.3
PA-DRL	**91.7**	**96.7**	**83.3**	**91.7**

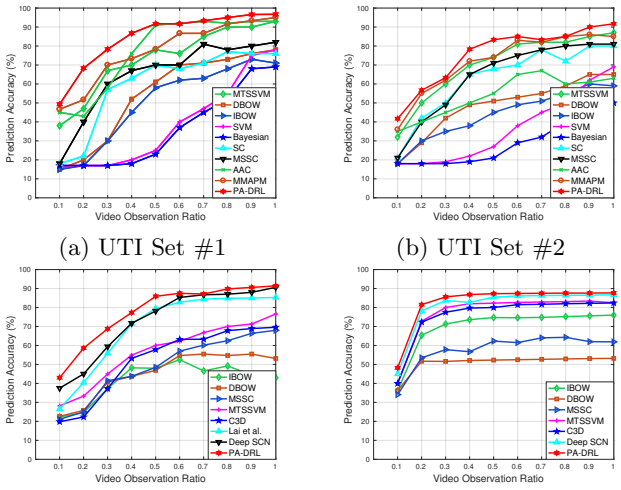
Table 2: The accuracy (%) of different methods on the BIT and UCF101 datasets.

Methods	BIT dataset		UCF101	
	OR=0.5	OR=1.0	OR=0.5	OR=1.0
IBoW [31]	49.2	43.0	74.6	76.0
DBoW [31]	46.9	53.1	53.2	53.2
MSSC [2]	48.4	68.0	62.6	61.9
MTSSVM [16]	60.0	76.6	82.3	82.5
Lai *et. al* [19]	79.4	85.3	-	-
Deep SCN [17]	78.1	90.6	85.5	86.7
C3D [39]	57.8	69.6	80.0	82.4
PA-DRL	**85.9**	**91.4**	**87.3**	**87.7**

Comparisons with the State-of-the-arts: From Table 1 and Table 2, we clearly see that PA-DRL achieves the performance of the state-of-the-art on three datasets. For the difference of three sets, we compared the results in three sets individually. On the UTI #1, the performance of our PA-DRL reached 91.7% and 96.7% at $OR = 0.5$ and $OR = 1.0$. At $OR = 0.5$, comparing with the AAC, our PA-DRL improved 3%. For the other approaches, our PA-DRL outperformed at least 3.5%. The result of our method demonstrated that our PA-DRL has the strong ability of representing at the half observation of actions on this set. Although our PA-DRL achieved the similar performance at $OR = 1.0$, we obviously outperformed than other methods on half observed videos.

The categories of actions in UTI #2 are the same as that of the UTI #1, but the variations of background in UTI #2 are larger, which makes it more difficult to predict the actions. Depending on the UTI #2 in Table 1, our PA-DRL achieved the best performance comparing with other methods. At $OR = 0.5$, PA-DRL raised 7.7% than AAC and improved 5% comparing with Lan *et. al* [20]. The large variation of background made the prediction difficult. Nevertheless, our proposed method performed best by extracting features from skeleton proposal, especially when the video was observed with a half.

The BIT dataset has more categories of actions and is more complex than the sets of the UTI. Nevertheless, we achieved the state-of-the-art with the accuracy of 85.9% and 91.4%. The complexity of actions reduced the predicting precision of other approaches. But for PA-DRL, the variance of actions would



(a) UTI Set #1 (b) UTI Set #2

(c) BIT-Interaction (d) UCF101

Fig. 5: **The accuracy curve of prediction.** The observation ratio changes from 0.1 to 1.0. The comparisons are presented on UTI #1, UTI #2, BIT and UCF101 dataset.

not change the order of joints on the skeleton, which could minimize the impact of complexity. Our PA-DRL outperformed Lai *et. al* [19] and Deep SCN [17], which are the leading approaches on BIT dataset. Because PA-DRL enhanced the action-related parts in features and made these parts much more attentional to the original features. The enhanced features highlighted the discriminative information and achieved PA-DRL to outperformed other methods.

UCF101 dataset is a dataset of action recognition, which is much larger than the previous two dataset and has collected from realistic videos. The complexity of actions makes the prediction more difficult. However, at $OR = 0.5$, PA-DRL outperformed 1.8% than Deep SCN [17], which obtained the state-of-the-art results on UCF101 for action prediction. Comparing with the action recognition method, C3D [39], PA-DRL raised the accuracy with 7.3%. The significant gap with demonstrated that when the action was observed incompletely, the method for full length action recognition had difficult to predict the action. While PA-DRL successfully predicted the actions on the incomplete action videos.

Fig. 5 illustrates the comparisons between PA-DRL and other approaches on three datasets. In this figure, the horizontal axis of the figure corresponds to the observation ratio, and the vertical axis represents the average prediction accuracy. Our PA-DRL outperformed the other methods on the BIT dataset.

From Fig. 5(a), we see that the prediction curve of PA-DRL increased rapidly from $OR = 0.3$ to $OR = 0.5$, and became stable since $OR = 0.7$. The methods of the SVM [31] and the Bayesian [31] relayed on the complete information of actions and achieved a good performance at $OR = 0.9$. The method of MTSSVM

[16] quickly increased at low observation ratio, especially at $OR = 0.3$. But our PA-DRL outperformed the MTSSVM since $OR = 0.4$. Comparing with the method of MMAPM [14], our PA-DRL obtained a better performance at $OR = 0.5$ and $OR = 0.7$ and reached the comparable results with the MMAPM. The result demonstrates that our proposed PA-DRL has the strong ability for representing the full length actions.

From Fig. 5(b), we see that our proposed method performed the best at $OR = 0.5$ and $OR = 1.0$. The repaid increasing from $OR = 0.3$ and $OR = 0.5$ indicated that our method effectively captured the evolution of actions with the increasing of observation ratio. The MTSSVM is benefited from using histogram features of both local and global information in temporal domain. The MMAPM used the multi-temporal scale to model the ongoing actions. However, our PA-DRL just used the global information with temporal pooling. The comparable results demonstrates that our PA-DRL exploited the structural information of the human body effectively and enhanced the discriminative power of features.

From Fig. 5(c), we see that the prediction of PA-DRL performed the state-of-the-art on the BIT dataset. The Lai *et. al* [19] and Deep SCN [17] are leading approaches on the BIT dataset, which do not mine the structural information of human. However, our PA-DRL achieved a higher performance with exploiting the structural information and mining the saliency information of human. The high performance with little observation of the whole video indicated that our PA-DRL could predict the activity at the early stage.

From Fig. 5(d), PA-DRL performed an accuracy with 81.5% even at $OR = 0.2$, which the observed video was just the beginning of actions. By activating the action-related parts, PA-DRL precisely predicted the direction of action evolution with a few frames of the beginning part of videos. Comparing with C3D, PA-DRL performed well with observation ratio higher than 0.6. Because the method failed to reduce the disturbance of action-unrelated parts in features. PA-DRL enhanced the action-related parts and made the predicted direction of the action evolution close to the actual direction.

Analysis of Different Components: To analyze the effectiveness of PA-DRL, we took the experiments comparing PA-DRL with skeleton feature without local patch feature and skeleton proposal based feature without deep reinforcement learning (DRL). Table 3 illustrates the results. In our experiments, skeleton feature without local patch feature denoted that the feature used for predicting action was the just the skeleton feature, which was composed of the position of skeleton joints. Skeleton proposal based feature without DRL utilized the feature of local patch around skeleton joints without activating process.

As can be seen in Table 3, PA-DRL effectively exploited the relationship between the part in feature and actions and used the action-related parts to predict the direction of action evolution. The setting of using skeleton feature obtained the worst performance on three sets, which indicated that without using the apparent information could substantially decrease the precision of prediction. The performance of skeleton proposal based feature was higher than that of skeleton feature, which indicated that the structural information and apparent

Table 3: The accuracy (%) of action prediction with different settings for PA-DRL on the UTI #1 , the UTI #2 and BIT dataset.

Different settings	Dataset	OR=0.5	OR=1.0
Skeleton feature without local patch feature	UTI #1	69.9	73.3
Skeleton proposal based feature without RL	UTI #1	76.7	91.7
PA-DRL	UTI #1	91.7	96.7
Skeleton feature without local patch feature	UTI #2	66.7	70.0
Skeleton proposal based feature without RL	UTI #2	70.0	86.7
PA-DRL	UTI #2	83.3	91.7
Skeleton feature without local patch feature	BIT	62.3	75.7
Skeleton proposal based feature without RL	BIT	68.6	87.5
PA-DRL	BIT	85.9	91.4

information were complementary for representing actions. The comparisons between skeleton proposal based feature and PA-DRL showed that there was a significant gap. The improvement demonstrated that the action-unrelated parts in the feature limited the representation ability, while our PA-DRL effectively exploited the relationship between feature and actions for representation.

From $OR = 0.5$, we see that the observed video did not contain the sufficient information of the action. Treating all parts of the feature equally reduced the ability of action-related parts to represent the action and relatively generated the disturbance by the noise. PA-DRL selected the action-related parts to inhibit the influence of noise and performed a significant improvement. At $OR = 1.0$, the feature with the complete information of the action reduced the disturbance of noise, which made the skeleton proposal based feature achieve a good performance. However, PA-DRL still outperformed skeleton proposal based feature by precisely predicting the direction of action evolution.

5 Conclusion

In this paper, we have proposed a part-activated deep reinforcement learning (PA-DRL) method for action prediction. The aim of our proposed PA-DRL is to learn the policy of activating action-related parts. Our PA-DRL exploits the structural information through extracting features by the skeleton proposal and mines the related information of human body for the ongoing actions. Experimental results on UTI, BIT and UCF101 dataset have been presented to demonstrate the effectiveness of the propose method for action prediction.

Acknowledgment

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFA0700802, in part by the National Natural Science Foundation of China under Grant 61672306, Grant U1713214, Grant 61572271, Grant 91746107, in part by the Shenzhen Fundamental Research Fund (Subject Arrangement) under Grant JCYJ20170412170602564 and in part by the Natural Science Foundation of Tianjin under Grant 16JCYBJC15900.

References

1. Caicedo, J.C., Lazebnik, S.: Active object localization with deep reinforcement learning. In: ICCV. pp. 2488–2496 (2015)
2. Cao, Y., Barrett, D., Barbu, A., Narayanaswamy, S., Yu, H., Michaux, A., Lin, Y., Dickinson, S., Mark Siskind, J., Wang, S.: Recognize human activities from partially observed videos. In: CVPR. pp. 2658–2665 (2013)
3. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR. pp. 7291–7299 (2017)
4. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR. pp. 6299–6308 (2017)
5. Du, Y., Wang, W., Wang, L.: Hierarchical recurrent neural network for skeleton based action recognition. In: CVPR. pp. 1110–1118 (2015)
6. Goodrich, B., Arel, I.: Reinforcement learning based visual attention with application to face detection. In: CVPRW. pp. 19–24 (2012)
7. Hinami, R., Mei, T., Satoh, S.: Joint detection and recounting of abnormal events by learning deep generic knowledge. In: ICCV. pp. 3619–3627 (2017)
8. Hu, J.F., Zheng, W.S., Ma, L., Wang, G., Lai, J.: Real-time rgb-d activity prediction by soft regression. In: ECCV. pp. 280–296 (2016)
9. Huang, C., Lucey, S., Ramanan, D.: Learning policies for adaptive tracking with deep feature cascades. In: ICCV. pp. 105–114 (2017)
10. Jie, Z., Liang, X., Feng, J., Jin, X., Lu, W., Yan, S.: Tree-structured reinforcement learning for sequential object localization. In: NIPS. pp. 127–135 (2016)
11. Karayev, S., Baumgartner, T., Fritz, M., Darrell, T.: Timely object recognition. In: NIPS. pp. 890–898 (2012)
12. Kingma, D.P., Ba, J.: Adam:a method for stochastic optimization. In: ICLR (2015)
13. Kong, X., Xin, B., Wang, Y., Hua, G.: Collaborative deep reinforcement learning for joint object search. In: CVPR. pp. 1695–1704 (2017)
14. Kong, Y., Fu, Y.: Max-margin action prediction machine. TPAMI **38**(9), 1844–1858 (2016)
15. Kong, Y., Jia, Y., Fu, Y.: Learning human interaction by interactive phrases. In: ECCV. pp. 300–313 (2012)
16. Kong, Y., Kit, D., Fu, Y.: A discriminative model with multiple temporal scales for action prediction. In: ECCV. pp. 596–611 (2014)
17. Kong, Y., Tao, Z., Fu, Y.: Deep sequential context networks for action prediction. In: CVPR. pp. 1473–1481 (2017)
18. Krull, A., Brachmann, E., Nowozin, S., Michel, F., Shotton, J., Rother, C.: Poseagent: Budget-constrained 6d object pose estimation via reinforcement learning. In: CVPR. pp. 6702–6710 (2017)
19. Lai, S., Zheng, W.S., Hu, J.F., Zhang, J.: Global-local temporal saliency action prediction. TIP pp. 1–14 (2017)
20. Lan, T., Chen, T.C., Savarese, S.: A hierarchical representation for future action prediction. In: ECCV. pp. 689–704 (2014)
21. Liang, X., Lee, L., Dai, W., Xing, E.P.: Dual motion gan for future-flow embedded video prediction. In: ICCV (2017)
22. Liang, X., Lee, L., Xing, E.P.: Deep variation-structured reinforcement learning for visual relationship and attribute detection. In: CVPR. pp. 848–857 (2017)
23. Littman, M.L.: Reinforcement learning improves behaviour from evaluative feedback. Nature **521**(7553), 445–451 (2015)

24. Ma, S., Sigal, L., Sclaroff, S.: Learning activity progression in lstms for activity detection and early detection. In: CVPR. pp. 1942–1950 (2016)
25. Mahmud, T., Hasan, M., Roy-Chowdhury, A.K.: Joint prediction of activity labels and starting times in untrimmed videos. In: ICCV. pp. 5773–5782 (2017)
26. Martinez, J., Black, M.J., Romero, J.: On human motion prediction using recurrent neural networks. In: CVPR. pp. 2891–2900 (2017)
27. Mathe, S., Pirinen, A., Sminchisescu, C.: Reinforcement learning for visual object detection. In: CVPR. pp. 2894–2902 (2016)
28. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al.: Human-level control through deep reinforcement learning. Nature **518**(7540), 529–533 (2015)
29. Qi, S., Huang, S., Wei, P., Zhu, S.C.: Predicting human activities using stochastic grammar. In: ICCV. pp. 1164–1172 (2017)
30. Ren, Z., Wang, X., Zhang, N., Lv, X., Li, L.J.: Deep reinforcement learning-based image captioning with embedding reward. In: CVPR. pp. 290–298 (2017)
31. Ryoo, M.S.: Human activity prediction: Early recognition of ongoing activities from streaming videos. In: ICCV. pp. 1036–1043 (2011)
32. Ryoo, M.S., Aggarwal, J.: Ut-interaction dataset, icpr contest on semantic description of human activities (sdha). In: ICPR. vol. 2, p. 4 (2010)
33. Ryoo, M.S., Matthies, L.: First-person activity recognition: What are they doing to me? In: CVPR. pp. 2730–2737 (2013)
34. Shi, Q., Cheng, L., Wang, L., Smola, A.: Human action segmentation and recognition using discriminative semi-markov models. IJCV **93**(1), 22–32 (2011)
35. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: NIPS. pp. 568–576 (2014)
36. Singh, G., Saha, S., Sapienza, M., Torr, P.H.S., Cuzzolin, F.: Online real-time multiple spatiotemporal action localisation and prediction. In: ICCV. pp. 3637–3646 (2017)
37. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
38. Supancic III, J.S., Ramanan, D.: Tracking as online decision-making: Learning a policy from streaming videos with reinforcement learning. In: ICCV. pp. 322–331 (2017)
39. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: ICCV. pp. 4489–4497 (2015)
40. Tudor Ionescu, R., Smeureanu, S., Alexe, B., Popescu, M.: Unmasking the abnormal events in video. In: ICCV. pp. 2895–2903 (2017)
41. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. In: CVPR. pp. 1290–1297 (2012)
42. Wang, L., Xiong, Y., Lin, D., Van Gool, L.: Untrimmednets for weakly supervised action recognition and detection. In: CVPR. pp. 4325–4334 (2017)
43. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: CVPR. pp. 4724–4732 (2016)
44. Xu, Z., Qing, L., Miao, J.: Activity auto-completion: predicting human activities from partial videos. In: ICCV. pp. 3191–3199 (2015)
45. Yoo, S.Y.J.C.Y., Yun, K., Choi, J.Y.: Action-decision networks for visual tracking with deep reinforcement learning. In: CVPR. pp. 2711–2720 (2017)
46. Zaki, H.F.M., Shafait, F., Mian, A.: Modeling sub-event dynamics in first-person action recognition. In: CVPR. pp. 7253–7262 (2017)
47. Zhuang, B., Liu, L., Shen, C., Reid, I.: Towards context-aware interaction recognition for visual relationship detection. In: ICCV. pp. 589–598 (2017)