EasyChair Preprint

№ 7042

A Discussion of Data Sampling Strategies for Early Action Prediction

Xiaofa Liu, Xiaoli Liu and Jianqin Yin

November 14, 2021

A Discussion of Data Sampling Strategies for Early Action Prediction

Xiaofa Liu[1], Xiaoli Liu[1], Jianqin Yin[1]

[1] Beijing University of Posts and Telecommunications, Beijing 100876, China
jqyin@bupt.edu.cn

Abstract. Action prediction aims to predict an ongoing activity from an incomplete video, which is an important branch of human activity analysis with the important application in a number of fields, such as security surveillance, human-machine interaction, automatic driving, etc. Due to time continuity, there are a large number of redundant frames in video action sequences, which often brings challenges such as low computational efficiency and noise for action prediction. Most of the existing works levarage dense sampling or sparse sampling for processing video frames and characterize actions. On the one hand, the dense sample-based method often introduces redundant noise for predictions, easily causing confusing of the action semantics. On the other hand, although sparse sample-based method can alleviate the problem of redundant noise to a certain extent, it ignores the impact of sampling rate on action representation. In this paper, we combine the two-stream network framework and the teacher-student network framework to build an action prediction model, and discuss the influence of action representation under different sampling rates for partial or full videos. In this way, we can select more appropriate frames for video representation and thus achieve more accurate action prediction. The method proposed in this paper has achieved the current state-of-the-art performance on the standard dataset, i.e., UCF101, which verifies the effectiveness of our method.

Keywords: Action prediction, Teacher-student framework, Data sampling.

1 Introduction

Early action prediction aims to recognize the semantic information of the action an ongoing video. With the development of deep network in image classification and video understanding, methods based deep learning has become the mainstream methods in the field of action prediction in recent years [1][2][3][4][5][6][7].

As shown in Figure 1, it is a challenging task to accurately and quickly recognize the semantics of current actions from an incomplete video, especially when the actions are performed at very early stages. At the same time, the motion information contained in partial videos with different observation rates is also very different, even for the same action. How to extract robust features from these incomplete videos while reducing the influence of redundant noise caused by the temporal continuity of video frames is very important to the problem of video action prediction.

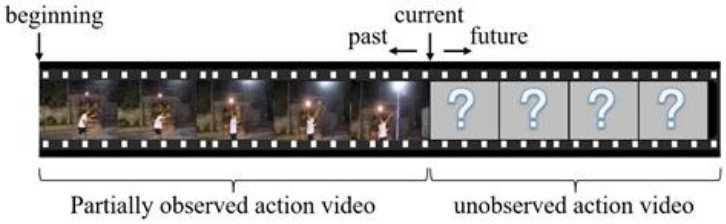

Fig. 1. Early action prediction, which predicts action label given a partially observed video.

Many works have been proposed for early action prediction. For instance, Kong et al. [1] extended marginalized stacked autoencoder (MSDA) to sequential data, which utilizes rich sequential context information to better capture the appearance evolution and temporal structure of the full action videos. To avoid the effect of noise caused by the background of the RGB frames as much as possible, Chen et al. [2] used the skeletal data to model and analyze actions. Liu et al. [3] also used the human skeletal sequence and introduced dilated convolutional network to model motion dynamics via a sliding window over the time axis. Gammulle et al. [5] proposed to use a GAN to generate future action descriptors and then classify them. Based on the idea of knowledge distillation, Wang et al. [6] used teacher network for action recognition to guide student network for prediction tasks, thus improving the accuracy of the prediction network. The above-mentioned works use sparse sampling strategies, i.e., a fixed number of frames are sampled for a video regardless of its length. This data sampling method alleviates the redundancy and noise problems caused by using all video frames to a certain extent, and fully utilizes the temporal information of the action sequence. But another problem is that different frame rates of sampling may affect the performance of early action prediction, which is ignored by the previous work. For example, key motion information is missing from few frames, while reductant noise from all video frames is disturbing the network.

In this work, we explore the impact of different data sampling rates on early human action prediction, aiming to provide a guiding significance for subsequent related works. Firstly, we sample different numbers of video frames for partial and full videos. Secondly, the pre-trained BN-Inception network on Kinetic-400 is used to extract features of partial and full videos, respectively. The teacher-student network framework proposed in [6] is used as the pipeline for early action prediction. We evaluate the performance on UCF101 dataset and obtain the current start-of-the-art (SOTA) performance, verifying the effectiveness of the proposed method. The experimental results show that it is unnecessary to use all frames for early action prediction, and different framerates have limited affect for predictions.

In summary, the main contributions of our work in this article are as follows:

- For data sampling, we discuss the efficiency of different data sampling rates on early action prediction, and provide a novel guidance significance for subsequent works.
- The proposed method is evaluated on the UCF101 dataset and achieves the state-of-the-art performance, verifying the influence of different sampling rates on early action prediction.

2 Related Work

In this part, we mainly discuss related works in the field of action prediction.

2.1 Action Recognition

Early action recognition mainly relied on manual features extracted from video (such as 3DHOG [8], SIFT [9], etc.) to model action appearance information and motion information. In recent years, as the progress of deep learning for a series of vision tasks, deep network has developed into the main methods for action recognition, which mainly includes Convolutional Neural Networks (CNN) [10][11][12] and Two-Stream networks [13][14], which achieves state-of-the-art recognition results on UCF101[15], Kinetics [16] and other datasets. Tran et al. [17] proposed the deep 3-dimensional convolutional networks (C3D) model, which used 3D ConvNets to model the spatio-temporal and motion information of video actions. The author verified that the linear classifier with C3D feature achieves the best effect in various video analysis tasks. However, the increasing depth of the network is limited due to the expensive computational cost and memory requirements of 3D ConvNets. Qiu et al. [17] proposed to decouple the 3D convolution into a 2D convolution for spatial modeling and a 1D convolution for temporal modeling. Therefore, the author built a Pseudo-3D Residual Networks (P3D) network to simulate 3D ConvNets to learn the spatio-temporal representation of videos, and verified the effectiveness and generalization of its spatio-temporal representation on five commonly used datasets. Different from 2D CNN and 3D CNN, Simonyan et al. [14] proposed the two-stream network that uses two parallel networks. One of which uses still images as input to obtain the appearance information of video actions, and the other uses multi-frame dense optical flow as input to obtain the motion information of the video actions. The two kinds of information are merged at the end to realize the final action classification. In order to use the information of the entire video without being limited by the length of time, Wang et al. [12] proposed Temporal Segment Networks (TSN) for long-term modeling. Firstly, the full video was divided into K segments, then each segment passes through the two-stream network to obtain the action representation and category scores. Finally, the two networks are merged to achieve video-level prediction. Because the spatial background of the video and the occurrence of actions often do not change synchronously, Feichtenhofer et al. [19] proposed a SlowFast network, which also uses two paths. One pathway is designed to capture semantic information that can be given by images or a few frames, which operates at low frame rates and slow refreshing speed. The other pathway is responsible for capturing rapidly changing motion by operating at fast refreshing speed and high temporal resolution. The two pathways are fused by lateral connections.

2.2 Action Prediction

For video action prediction, many methods based on deep learning have emerged in recent years. Kong et al. [1] proposed a Deep Sequential Context Networks (Deep-

SCN) for early action prediction. The author believed that the confidence of prediction increases with the number of observed frames. In [1], the author directly used the C3D features [17] generated by video frames through a structured support vector machines (SVM) to capture the time structure of human behavior. Chen et al. [2] extracted human skeleton points under the framework of deep reinforcement learning to characterize the human body structure. Their proposed method activated the action-related parts of the feature to capture action-related information and suppress the influence of noise. To solve the different duration of different actions, Liu et.al. [3] proposed a time scale selection network Scale Selection Network (SSNet), which adaptively selects the number of frames for prediction according to the duration of the action. In this way, the author can suppress the influence of noise. To make full use of the global information of the video, Wang et al. [6] proposed to distill some useful knowledge from the teacher model to facilitate the student prediction model. Although the above works [1][2][3][6] did not use all frames of the video, they have not discussed and explored the impact of different number of the videos for early action prediction. Therefore, in this paper, we discuss the influence of different sampling rates on action representation in action prediction.

3 Methodology

In this section, we introduce in detail how our manuscript performs data sampling strategies and feature extraction. First, we introduce the overall framework of the network. Then we introduce the sampling method and feature extraction method in our framework.

As shown in Figure 2, in our work, we divide a full video into N ($N = 10$) sub-segments of equal length. The first n sub-segments are defined as the progress level n with an observation rate of n/N. We use x_n to represent the feature of the sub-segment of the progress level n.

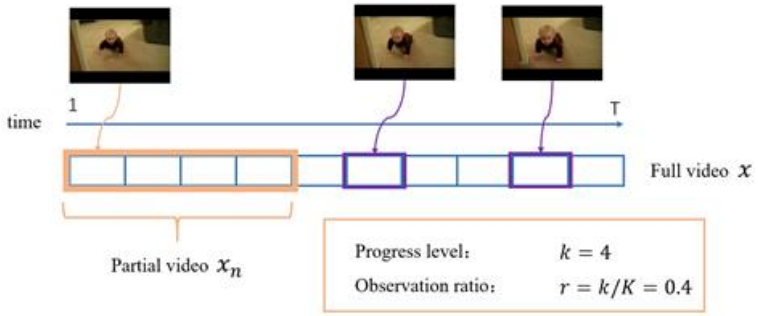

Fig. 2. **Definition of some concepts.** Taking the progress level of 4 as an example, this figure shows the division of video, the definition of the progress level, and the representation of features in motion prediction.

3.1 Overview

In this work, we use the teacher-student network framework proposed in [6] as the basic model. The overall network framework is shown in Figure 3.

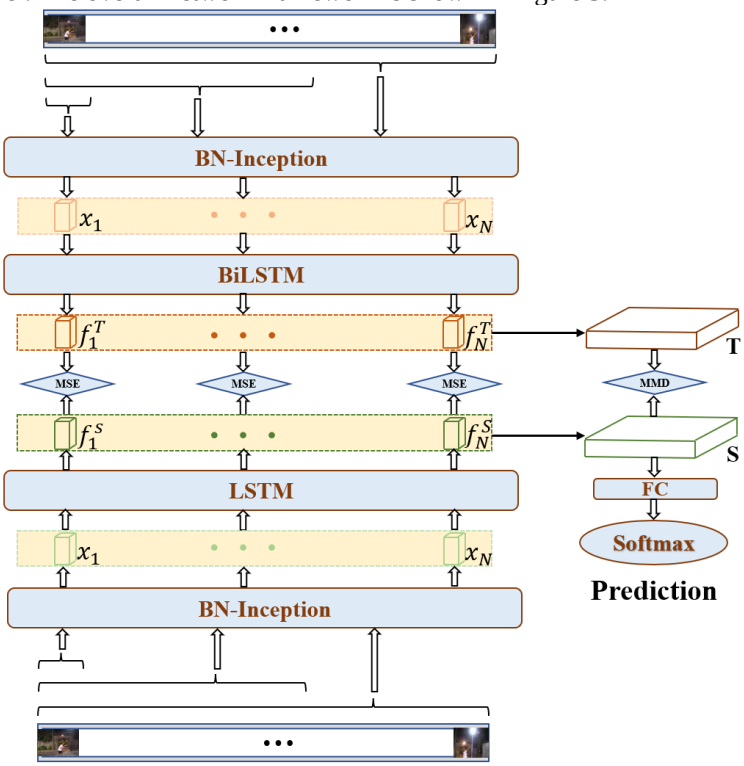

Fig. 3. Overall framework. The overall structure of the network based on the framework of the teacher-student network in [6].

Given the currently observed partial action video, our goal is to predict the action semantic label y of the video.

The teacher network can be defined as:

$$F_T = T\big(C(X)\big) \tag{1}$$

Where X represents the frame after sampling the fragments of different progress levels. $C(\cdot)$ represents the feature extraction through the convolutional neural network [20]. The feature representation of different progress levels generated after the convolutional neural network are recorded as $x_i,\ i = (1, 2, \dots, N)$. $T(\cdot)$ represents the teacher network [6]. After passing through the teacher network, the feature representation under different progress levels are recorded as f_i^T, and the feature representation under all progress levels are recorded as $T, T = \{f_1^T, f_2^T, \dots, f_N^T\}$.

The student network used for prediction can be defined as:

$$F_S = S\big(C(X)\big) \tag{2}$$

$$y = V\big(L(F_S)\big) \tag{3}$$

Where X and $C(\cdot)$ indicate the same as the teacher network in formula (1). $S(\cdot)$ represents the student network [6]. After passing through the student network, the feature representation under different progress levels are recorded as f_i^S, and the feature representation under all progress levels are recorded as S, $S = \{f_1^S, f_2^S, \dots, f_N^S\}$. The feature F_S obtained by the student network passes through the linear layer $L(\cdot)$ and the final Softmax layer $V(\cdot)$ to finally obtain the predicted label y.

The teacher-student network is the same as [6]. In the following sections we will further describe the feature extraction and data sampling strategies.

3.2 Feature Processing

To explore the number of frames of partial and full videos for early action prediction, we represent the human action under different sampling rates for early action prediction. Our detailed implementation for feature extraction is as follows. Firstly, we extract the dense optical flow characterization A from the full video. Then we use a sliding window with a size of 5 and a stride of 1 to sample a number of frames on A, and the sampled frames are used as the input of BN-Inception [20] for feature extraction. Finally, we obtain the feature representation of the full video, denoting by B.

3.3 Date Sampling Strategies

To obtain the feature representation of different progress levels for prediction, we obtain the feature representation for different partial videos from the feature representation B of the full video. Finally, we use different sampling rates to sample the features of the partial videos or full videos, and the final features are obtained by mean pooling operation, denoting by x_i, $i = (1, 2, \dots, N)$, where x_i is a one-dimensional feature vector with a size of 1024.

4 Experiments

We test our proposed method on a benchmark dataset, i.e., UCF-101[15]. Below, we will analyze the experimental details and results.

4.1 Implementation Details

We use the same experimental settings as in [6] on the RGB dataset UCF101. To generate feature representation for teacher and student network learning, we use the pre-trained BN-Inception network on Kinetic-400 [16] to extract the features of the partial video and the full video, respectively. For the partial video and the full video, we use a stride length of 5 to sample L frames, ranging from 10 to 60, to form the video representation.

4.2 Experiments on the UCF-101 Dataset

The UCF101 dataset [15] contains 101 categories and a total of 13,320 videos. The video duration is about a few seconds of length.

The detailed experimental results are shown in Table 1. Compared with the baselines that use all frames, the accuracy of our method is the best regardless of the sampling rates. The experimental results show that it is unnecessary to use all frames for early action recognition. Proves that using all frames may cause redundant noise for accurate prediction. As shown in Table 1, the performance of different sampling rates is similar. And the sampling rate ranging from 15 to 20 works the best. When the number of sampling frames continues to increase, it will introduce more irrelevant information for the action, which easily cause misclassification, especially using all video frames for the prediction.

Table 1. Prediction results (%) on the UCF101 set.

Observation ratio	10%	30%	50%	70%	100%	Mean
Baseline	75.34	90.28	93.26	93.92	94.87	91.04
L=10	86.39	91.25	93.45	94.38	95.30	92.68
L=15	86.42	91.61	93.56	94.79	95.36	92.88
L=20	86.37	91.39	93.59	94.41	95.25	92.75
L=25	86.26	91.25	93.59	94.49	95.38	92.79
L=30	86.12	91.50	93.40	94.43	95.22	92.67
L=35	86.34	91.55	93.51	94.68	95.46	92.79
L=40	86.04	91.47	93.35	94.46	95.33	92.68
L=45	86.50	91.31	93.56	94.51	95.19	92.68
L=50	86.58	91.44	93.35	94.51	95.46	92.75
L=55	86.18	91.44	93.37	94.51	95.46	92.75
L=60	86.01	91.44	93.43	94.35	95.33	92.67

5 Conclusion

In this paper we discuss the redundancy of videos in the field of video action prediction, and verified the influence of action representation under different sampling rates on the accuracy of action prediction. We have empirically shown that it is unnecessary to use all frames of the video for early action prediction, and different sampling rates of the videos show similar performance, but 15~20 frames are the more proper sampling rates. We hope that the discussion in this paper will provide guiding significance for future work in the field of video action prediction. How to make full use of the limited information while reducing the impact of video redundancy and noise has further research significance for action representation and predictive performance.

6 Future Work

If necessary, to obtain detailed experimental support, we plan to use more backbones for verification on more datasets.

Acknowledgement

This work was supported partly by the National Natural Science Foundation of China (Grant No. 62173045, 61673192), and partly by the Fundamental Research Funds for the Central Universities (Grant No. 2020XD-A04-2).

References

1. Kong, Y., Tao, Z., Fu, Y, Deep sequential context networks for action prediction, In: CVPR, 2017, pp. 1473–1481.
2. Chen, L., Lu, J., Song, Z., & Zhou, J, Part-activated deep reinforcement learning for action prediction. In: ECCV, 2018, pp. 421-436.
3. Liu, J., Shahroudy, A., Wang, G., Duan, L. Y., & Kot, A. C, SSNet: scale selection network for online 3D action prediction, In: CVPR, 2018, pp. 8349-8358.
4. Zhao, H., & Wildes, R. P, Spatiotemporal feature residual propagation for action prediction, In: ICCV, 2019, pp. 7003-7012.
5. Gammulle, H., Denman, S., Sridharan, S., & Fookes, C, Predicting the future: A jointly learnt model for action anticipation, In: ICCV, 2019, pp. 5562-5571.
6. Wang, X., Hu, J. F., Lai, J. H., Zhang, J., & Zheng, W. S, Progressive teacher-student learning for early action prediction, In: CVPR, 2019, pp. 3556-3565.
7. Scarafoni, D., Essa, I., & Ploetz, T, PLAN-B: Predicting Likely Alternative Next Best Sequences for Action Prediction, 2021, arXiv preprint arXiv:2103.15987.
8. Alexander Klaser, Marcin Marszałek, and Cordelia Schmid, A spatio-temporal descriptor based on 3d-gradients, In: British Machine Vision Conference, 2008, pp. 275–1.
9. Paul Scovanner, Saad Ali, and Mubarak Shah, A 3-dimensional sift descriptor and its application to actionrecognition, In: ACM international conference on Multimedia, 2007, pp. 357–360.
10. Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset, In: CVPR, 2017, pp. 4724–4733.
11. Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh, Learning spatio-temporal features with 3d residual networks for action recognition, In: ICCV, 2017, pp. 4.
12. Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool, Temporal segment networks: Towards good practices for deep action recognition, In ECCV, 2016, pp. 20–36.
13. Yi Zhu, Zhen-Zhong Lan, Shawn D. Newsam, and Alexander G. Hauptmann, Hidden two-stream convolutional networks for action recognition, CoRR, 2017, abs/1704.00389.
14. Simonyan, K., Zisserman, A, Two-stream convolutional networks for action recognition in videos, In: NIPS, 2014, pp. 568-576.
15. Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah, Ucf101: A dataset of 101 human actions classes from videos in the wild, CoRR, 2012, abs/1212.0402.
16. Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman, The kinetics human action video dataset, CoRR, 2017, arXiv

preprint arXiv:1705.06950.
17. Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, Learning spatiotemporal features with 3d convolutional networks, In: ICCV, 2015, pp. 4489–4497.
18. Qiu Z, Yao T, Mei T, Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks, In: ICCV, 2017, pp. 5533-5541.
19. Feichtenhofer C, Fan H, Malik J, et al, SlowFast Networks for Video Recognition, In: ICCV, 2019, pp. 6202-6211.
20. Ioffe, S., & Szegedy, C., Batch normalization: Accelerating deep network training by reducing internal covariate shift, In: International conference on machine learning. PMLR, 2015. pp. 448-456.