Early Action Prediction and Analysis from Frame Glimpses in Video

Xin Ye, Zige Huang

1 Introduction

FOR active human assistant robots, it is very important for them to predict human intentions so that they can help human in time. To be specific, if human is doing some kind of activities and if the human assistant robot can predict the next action human will do, the robot can generate correspondent plannings and then help human finish their tasks. Due to this motivation, we are going to do the early action prediction from vision input, since it is much easier for robots to have a camera than any other sensors.

We will take a video sequence as an input, and then predict the next action label. We plan to use two different methods to do the experiments. Firstly, we will use CNN-RNN-based model to do the prediction and use backpropagation algorithm to train the model. Here, CNN model is used to extract image features while RNN model is used to capture time series. Then we will try to use reinforcement leaning, and follow the work from [10] to introduce a reward function to learn when is the best time to emit prediction. Finally, we will compare and analysis these two methods from different aspects, like which one can achieve a higher accuracy while observing only a fraction of the frames.

We will evaluate our model on the THUMOS'14 and ActivityNet datasets. The unconstrained and untrimmed setting of the THUMOS'14 and ActivityNet action detection dataset can bring us more possibilities for validation. And we will also do comparison from different aspect like the accuracy between our computational method and data from human observers (ground truth). For different kinds of action, we plan to list average errors and analyze what is the reason to cause this difference. This kind of data will give our prediction a higher accuracy.

2 Related Work

Analysis of video learning and the early events prediction have a very long history and many different topics. We will focus reviewing recent works about segmentation algorithm, early actions detection and temporal behavior prediction works, as well as action classification as an online process.

In the aspect of segmentation algorithm, [1] introduces a new algorithm for object segmentation in complex environment. It solves the task of challenging the segmentation of unknown objects, especially the 3D point clouds of non-convex objects.

For early actions detection works, a fast target detection framework for real-world robot applications is presented in [2]. [4] shows us a method to do real-time early events detection on the big social data. [10] introduces an end to end method to learn and detect the video for a given data which also use the reward mechanism to solve the variable-sized outputs problem. Those papers above give us a very good thinking for our works about how to detect the given action in the predicted time range.

For the temporal behavior prediction, time conditional random field model is used to infer the expected human activities [5]. And [6] predicts the future trajectory of surveillance video by using concept detectors. In addition, Yuan [7] combine dense trajectory for temporal motion detection with CNN frame-level features and sound characteristics of sliding window in the framework. The time of the child component is detected by setting syntax on different complex actions in [8].

In the direction of action classification, [3] enjoys great popularity for their performance and their way to classify and extract the related features simultaneously.

There are also many works very related to our work like Oneata [8] and Wang [9] which use the same unconstrained and untrimmed datasets as our work will do. The analysis of the entire data set in their work is of great value to us.

Above papers provide us broad thinking for forecasting problems. They try to solve this problem from different aspect and have special achievements in temporal prediction. Those papers are very helpful for us to do action prediction based on neural network.

3 Time Arrangement

Here is our time line to finish this project.

From Feb.25 to Mar.6:

Research and study related work about early action detection as well as reinforcement learning (Xin and Zige).

From Mar.7 to Apr.3:

Collect public dataset, such as THUMOS'14 and ActivityNet (Xin and Zige);

Implement the first method, CNN-RNN-based method. (Zige)

Implement the second method by following the work from [10], which is the reinforcement learning based method. (Xin)

From Apr.4 to Apr.17:

Evaluate two methods according to the prediction results, and try to analysis the pros and cons of these two methods (Xin and Zige).

From Apr.18 to Apr.30:

Finish the report for this final project (Xin and Zige).

REFERENCES

[1] Aleksandrs Ecins, Cornelia Ferm¨uller, and Yiannis Aloimonos. Cluttered scene segmentation using the symmetry constraint. In ICRA, 2016 W.-K. Chen, *Linear Networks and Systems.* Belmont, Calif.: Wadsworth, pp. 123-135, 1993. (Book style)

[2] Luan, W., Yang, Y., Fermuller, C., & Baras, J. S. Fast task-specific target detection via graph based constraints representation and checking.2016

[3] H. Li, Y. Li, and F. Porikli. Deeptrack: Learning discriminative feature representations online for robust visual tracking. IEEE Transactions on Image Processing, 25(4):1834–1848, April 2016.

[4] Aitor Aldoma, Federico Tombari, Luigi Di Stefano, and Markus Vincze. A global hypotheses verification method for 3d object recognition. In Proceedings of the 12th European Conference on Computer Vision - Volume Part III, ECCV'12, pages 511–524, Berlin, Heidelberg, 2012. Springer-Verlag.

[5] Koppula H, Saxena A. Anticipating human activities using object affordances for reactive robotic response. IEEETransactionson pattern Analysis and Machine Intelligence 38(1) ,2016.

[6] Kitani KM, Ziebart BD, Bagnell JA, Hebert M Activity forecasting. In: European Conference on Computer Vision (ECCV)),2012.

[7] J. Yuan, Y. Pei, B. Ni, P. Moulin, and A. Kassim. Adsc submission at thumos challenge 2015.

[8] H. Pirsiavash and D. Ramanan. Parsing videos of actions with segmental grammars. In Computer Vision and Pattern Recognition (CVPR), 2014.

[9] D. Oneata, J. Verbeek, and C. Schmid. The lear submission at thumos 2014. 2014.

[10] Yeung, Serena, et al. "End-to-end learning of action detection from frame glimpses in videos." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.