Long-term Action Anticipation: A Quick Survey

Zeyun Zhong

Vision and Fusion Laboratory
Institute for Anthropomatics
Karlsruhe Institute of Technology (KIT), Germany
zeyun.zhong@kit.edu

Abstract

The ability to anticipate possible human actions in the distant future is of fundamental interest for a wide range of applications, including autonomous driving, surveillance, and human-robot interaction. Consequently, various methods have been presented for action anticipation in recent years, with deep learning-based approaches being particularly popular. In this work, we give a short overview of the recent advances of long-term action anticipation algorithms.

1 Introduction

In the last years, we have seen a tremendous progress in the capabilities of computer systems to classify and segment activities in videos. These systems, however, analyze the past or in the case of real-time systems the present with a delay of a few milliseconds. For applications, where a moving system has to react or interact with humans, this is insufficient. For instance, to be able to offer a hand at the right time or to generate proactive dialog to provide more natural interactions, collaborative robots that work closely with humans have to anticipate the activities of a human in the future. Compared to human action recognition and early action recognition, where entire or part of action

segments are observable, action anticipation aims to predict future action without observing any part of it, as displayed in Figure 1.1.

As the anticipation results are just assumptions, this tends to be significantly more challenging than traditional action recognition, which performs well with todays well-honed discriminative models [7, 17]. Consistent with action recognition, anticipation approaches start with prediction on only one single video frame [28] and tend to use longer temporal context [24, 29] in recent years. Apart from using a long action history, many approaches attempt to leverage several modalities other than just the raw video frames, such as the motion information and objects contained in the scene, to further improve the predictive ability.

While many recent works anticipate activities only for a very short time horizon of a few seconds [9, 8], there is a parallel line of work [6] which addresses the problem of anticipating all activities that will be happening within a time horizon of up to several minutes, which is particularly interesting for robot systems that require certain time to react and plan the future tasks.

In spite of the enormous amount of research conducted in this area, the problem is still challenging due to the fundamental challenges inherent to the task such as the multi-modal distribution of future action candidates, especially for the scenario where we are going to predict far into the future (long-term anticipation). As action recognition is usually a fundamental sub-component of an anticipation system, the challenges of action recognition [14] are also included, such as the tremendous intra-class variance among the activities, huge spatio-temporal scale variation, target motion variations, etc. Moreover, low image resolution, object occlusion, illumination change and viewpoint change further aggravate these challenges.

Although classical learning approaches, such as Conditional Random Fields (CRFs) [15], Markov models [23], and other statistical methods [19, 22], have been widely used in the literature, we put our focus on deep learning techniques and how they have been extended or applied to daily-living action anticipation, leaving the classical approaches outside the scope of the present review. In this context, the terms action anticipation, action prediction, and action forecasting are used interchangeably.

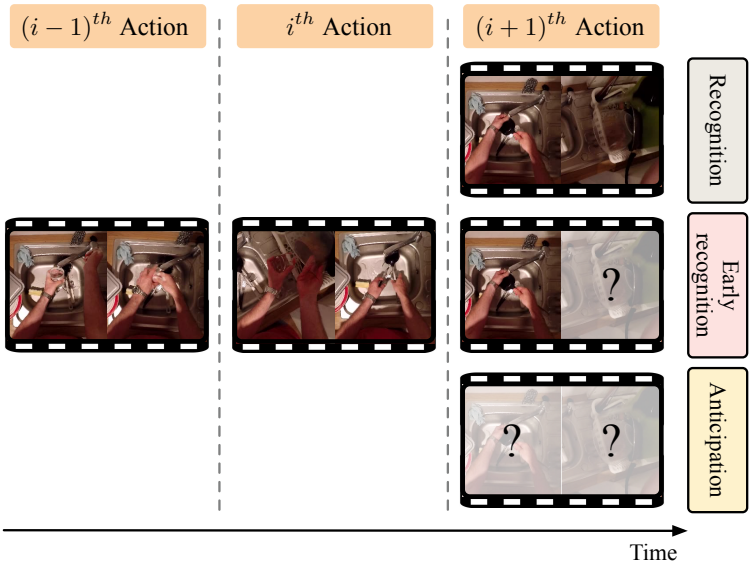


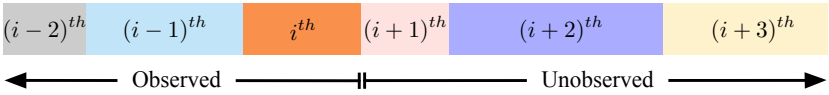
Figure 1.1: The action anticipation task aims to anticipate future action(s) before it happens, whereas action recognition and early action recognition require the observation of complete and partial actions, respectively.

This survey is structured as follows. In Section 2, we describe both short-term and long-term anticipation tasks which are commonly used in the literature, so that the reader can better distinguish between them. In Section 3, we introduce the current approaches that address the long-term anticipation task and discuss their limitations. Finally, we conclude this survey in Section 4.

2 Problem Statement

Based on the prediction time horizon, action anticipation approaches can be grouped into two categories: short-term anticipation approaches and long-term anticipation approaches. While short-term approaches predict a single action a

(a) Short-term anticipation.



(b) Long-term anticipation.

Figure 2.1: Category of the action anticipation task. While the short-term anticipation aims at predicting a single future action, long-term task aims to predict a sequence of the following actions.

few seconds into the future, long-term approaches predict a sequence of future actions with their durations up to several minutes into the future. In the following sections, we show the detailed task definition of both categories usually used in the literature.

2.1 Short-term anticipation

Most short-term anticipation approaches follow the setup defined in [28, 4, 5]. As illustrated in Figure 2.1(a), the task aims to predict a future action by observing a video segment of length τ_o. The observation segment is τ_a seconds preceding the action, i.e., from time $\tau_s - (\tau_a + \tau_o)$ to $\tau_s - \tau_a$, where τ_a denotes the "anticipation time", i.e., how many seconds in advance actions are to be anticipated. The anticipation time τ_a is usually fixed for each dataset, whereas the length of the observation segment is typically dependent on the individual method. Methods in this category typically use synchronous data to perform the anticipation task, meaning that the input to the model is a sequence of frames that have the same temporal spacing before the action [9, 8].

Some work [18, 21] attempts to predict the starting time of the next action as well. As this task involves the duration of each action, these approaches usually use asynchronous data as input to the model, containing a sequence of action categories and inter-arrival times. The inter-arrival time is defined as the difference between the starting time of last and the current action. With the predicted inter-arrival time, the starting time of the next action can be easily deduced.

2.2 Long-term anticipation

There is a parallel line of research addressing the long-tern anticipation task, which is proposed in [6]. The goal is to anticipate the category and the duration of future actions for a given time horizon, which can take up to several minutes, as illustrated in Figure 2.1(b). Long-term approaches typically take a sequence of observed action categories and their durations to predict another sequence of actions and durations [6, 1, 31].

3 Long-term Anticipation Approaches

3.1 Methods

Farha et al. [6] first introduced the long-term action anticipation task and proposed two models to tackle the task. One is based on an RNN model, which outputs the remaining length of the current action, the next action class and its length, as shown in Figure 3.1. The long-term prediction is conducted recursively, i.e., observations are combined with the current prediction to produce the next prediction. Another method is based on a CNN model, which outputs a sequence of future actions in a form of a matrix in one single step. Considering the limitations of these two methods, i.e., the RNN model is time-consuming and suffers from error accumulation and the CNN model introduces many parameters when predicting long sequences, Ke et al. [12] proposed a method to explicitly address these issues. They chose to condition on a time variable representing the prediction horizon. Specifically, they transformed the prediction time horizon

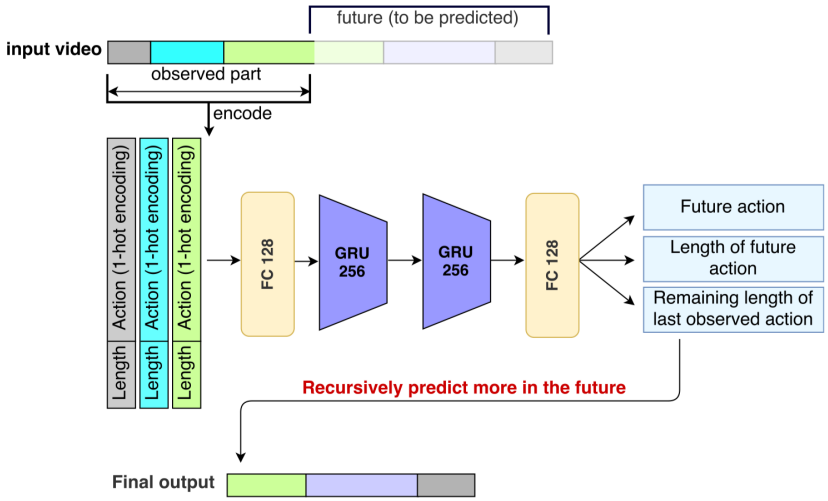


Figure 3.1: Architecture of the RNN system [6]. The input is a sequence of (length, 1-hot class encoding)-tuples. The network predicts the remaining length of the last observed action and the label and length of the next action. Appending the predicted result to the original input, the next action segment can be predicted. Figure is taken from [6].

to a time representation, and concatenated it with the original inputs forming time-conditioned observations. Their model is therefore capable of anticipating a future action at arbitrary and variable time horizons in a one-shot fashion. Additionally, they introduced a time-conditioned skip connection between the last observed action and the initial anticipation based on the intuition that the last action of the observation is generally relevant to the future actions.

Inspired by [12], Gong et al. [10] proposed an encoder-decoder structure based on transformer architecture [27, 2], which effectively captures long-term relations over the whole sequence of actions. The encoder learns to capture fine-grained long-range temporal relations between the observed frames from the past, while the decoder learns a sequence of future action queries, capturing global relations between upcoming actions in the future along with the observed features from the encoder. Because of the proposed parallel decoding, the model is able to

make more accurate and faster inference without potential error accumulations caused by autoregressive decoding. However, the number of predictable future actions is also limited to the number of action queries used in the training process, which might need to be adapted, if the model is applied for other datasets.

Predicting future is inherently multi-modal. Given an observed video segment containing an ongoing action, multiple actions could be possible to be the next action following the observed one. This uncertainty becomes even larger if we are going to predict far into the future. Therefore, it may be beneficial to model the underlying uncertainty, allowing to capture different possible future actions. However, in most approaches, action prediction is taken as a classification problem and optimized under cross-entropy loss, suffering from overly high resemblance to dominant ground truth, while suppressing other reasonable possibilities [3]. Moreover, approaches that are optimized with mean square error tend to produce the mean of the modes [28, 20]. To this end, some approaches are proposed to tackle the uncertainty in the future predictions, which are described below.

Farha and Gall [1] introduced a framework that predicts all subsequent actions and corresponding durations in a stochastic manner. In their framework, an action model similar to the one proposed in [6] (shown in Figure 3.1) and a time model are trained to predict the probability distribution of the future action label and duration, respectively. While action labels are taken as classifications and optimized under cross-entropy (CE) loss, durations are taken as real-valued variables which are modeled with a Gaussian distribution and optimized with the negative log likelihood (NLL). At test time, future action label and its duration are sampled from the learned distributions. Long-term predictions are achieved by feeding the predicted action segment to the model recursively.

Zhao and Wildes [31] proposed Conditional Adversarial Generative Networks to address the underlying uncertainty when predicting future action sequence. More specifically, different from many works that operate with continuous time variable [6, 1, 12, 10], they treated both action labels and time as discrete data which are formated as one-hot vectors. These vectors are first projected to higher dimension continuous spaces and concatenated, and then fed to a *seq2seq* generator [26] to compute logits of future action labels and their corresponding time. To obtain differentiable sampling to generate future sequences with

both quality and diversity during training, the Gumbel-Softmax relaxation technique [11] that mimics one-hot vectors from categorical distributions and a normalized distance regularizer [30] that encourages diversity are adopted. A ConvNet classifier is used as the discriminator to allow to train the generator adversarially.

Mehrasa et al. [21] proposed using a recurrent variational auto-encoder (VAE [13]) to capture the distribution over the times and categories of action sequences. To overcome the problem that a fixed prior distribution of the latent variable (usually $\mathcal{N}(0, I)$ in VAE models) may ignore temporal dependencies present between actions, authors learned a prior that varies across time. At test time, a latent code is sampled from the learned prior distribution, based on which the probability distributions of the action class and the corresponding time are inferred.

3.2 Limitations

Despite the impressive performance on the standard benchmarks [25, 16], current approaches have several limitations, which are described below.

Limited representativity of the evaluation datasets. The commonly used benchmark datasets for long-term anticipation, i.e., Breakfast [16] and 50Salads [25], contain only videos of a specific kitchen activity, which usually last several minutes. Since there is only one activity per video, i.e., either *preparing a breakfast* or *preparing a salad*, it is easier to predict the following actions than in the real-world scenarios, where a completely different action might occur next. Furthermore, since these videos are typically only several minutes long, the current setting may not be directly applicable for longer videos, especially for real-world applications. Moreover, these datasets do not contain any concurrent actions. However, actions in the real-world scenarios, such as *making a phone call* and *taking notes* may be performed simultaneously.

Difficult deployment of methods that incorporate uncertainty. Methods that incorporate uncertainty typically learn a joint distribution of all data samples. For evaluation, authors usually draw many samples from the learned distribution, and compute the average metric value of all drawn samples [1, 31], or select

the most frequent sample as the final result [21]. However, such an evaluation protocol requires multiple runs of the model, which is time-consuming and therefore difficult to deploy for real-time systems.

4 Conclusion

In this survey, we gave a short overview of the current approaches that are proposed to tackle the long-term action anticipation task. We analyzed different methods from two perspectives: research question each individual method addresses and method description. In the end, we also described the limitations of the current approaches. In conclusion, long-term action anticipation is an interesting and relatively new research topic, which attracts increasing attention in the community, and benefits many intelligent decision-making systems. While great strides have been made, there is still large room for improvement in action anticipation using deep learning techniques.

References

[1] Yazan Abu Farha and Juergen Gall. “Uncertainty-Aware Anticipation of Activities”. In: *ICCV Workshop*. 2019.

[2] Nicolas Carion et al. “End-to-end object detection with transformers”. In: *ECCV*. 2020, pp. 213–229.

[3] Bo Dai et al. “Towards diverse and natural image descriptions via a conditional gan”. In: *ICCV*. 2017, pp. 2970–2979.

[4] Dima Damen et al. “Scaling egocentric vision: The epic-kitchens dataset”. In: *ECCV*. 2018, pp. 720–736.

[5] Dima Damen et al. “The epic-kitchens dataset: Collection, challenges and baselines”. In: *TPAMI* 43.11 (2020), pp. 4125–4141.

[6] Yazan Abu Farha, Alexander Richard, and Juergen Gall. “When Will You Do What? - Anticipating Temporal Occurrences of Activities”. In: *CVPR*. 2018. arXiv: 1804.00892.

[7] Christoph Feichtenhofer et al. “Slowfast networks for video recognition”. In: *ICCV*. 2019, pp. 6202–6211.

[8] Antonino Furnari and Giovanni Farinella. “What Would You Expect? Anticipating Egocentric Actions With Rolling-Unrolling LSTMs and Modality Attention”. In: *ICCV*. 2019.

[9] Jiyang Gao, Zhenheng Yang, and Ram Nevatia. “RED: Reinforced Encoder-Decoder Networks for Action Anticipation”. In: *BMVC*. 2017.

[10] Dayoung Gong et al. “Future Transformer for Long-term Action Anticipation”. In: *CVPR*. 2022. arXiv: 2205.14022 [cs].

[11] Eric Jang, Shixiang Gu, and Ben Poole. “Categorical reparameterization with gumbel-softmax”. In: 2017.

[12] Qiuhong Ke, Mario Fritz, and Bernt Schiele. “Time-Conditioned Action Anticipation in One Shot”. In: *CVPR*. June 2019.

[13] Diederik P Kingma and Max Welling. “Auto-encoding variational bayes”. In: *ICLR*. 2014.

[14] Yu Kong and Yun Fu. “Human action recognition and prediction: A survey”. In: *IJCV* 130.5 (2022), pp. 1366–1401.

[15] Hema S. Koppula and Ashutosh Saxena. “Anticipating Human Activities Using Object Affordances for Reactive Robotic Response”. In: *TPAMI* 1 (2016).

[16] Hilde Kuehne, Ali Arslan, and Thomas Serre. “The Language of Actions: Recovering the Syntax and Semantics of Goal-Directed Human Activities”. In: *CVPR*. 2014.

[17] Ze Liu et al. “Video swin transformer”. In: *CVPR*. 2022, pp. 3202–3211.

[18] Tahmida Mahmud, Mahmudul Hasan, and Amit K Roy-Chowdhury. “Joint prediction of activity labels and starting times in untrimmed videos”. In: *ICCV*. 2017, pp. 5773–5782.

[19] Tahmida Mahmud et al. “A poisson process model for activity forecasting”. In: *ICIP*. IEEE. 2016, pp. 3339–3343.

[20] Michael Mathieu, Camille Couprie, and Yann LeCun. “Deep multi-scale video prediction beyond mean square error”. In: *ICLR*. 2016.

[21] Nazanin Mehrasa et al. “A Variational Auto-Encoder Model for Stochastic Point Processes”. In: *CVPR*. 2019.

[22] Siyuan Qi et al. “Predicting Human Activities Using Stochastic Grammar”. In: *ICCV*. 2017.

[23] Nicholas Rhinehart and Kris M Kitani. “First-person activity forecasting with online inverse reinforcement learning”. In: *ICCV*. 2017, pp. 3696–3705.

[24] Fadime Sener, Dipika Singhania, and Angela Yao. “Temporal Aggregate Representations for Long-Range Video Understanding”. In: *ECCV*. 2020. arXiv: 2006.00830.

[25] Sebastian Stein and Stephen J McKenna. “Combining embedded accelerometers with computer vision for recognizing food preparation activities”. In: *UbiComp*. 2013, pp. 729–738.

[26] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. “Sequence to sequence learning with neural networks”. In: *NeurIPS*. 2014.

[27] Ashish Vaswani et al. “Attention Is All You Need”. In: *NeurIPS*. 2017.

[28] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. “Anticipating visual representations from unlabeled video”. In: *CVPR*. 2016, pp. 98–106.

[29] Chao-Yuan Wu et al. “MeMViT: Memory-Augmented Multiscale Vision Transformer for Efficient Long-Term Video Recognition”. In: *CVPR*. 2022. arXiv: 2201.08383.

[30] Dingdong Yang et al. “Diversity-sensitive conditional generative adversarial networks”. In: *ICLR*. 2019.

[31] He Zhao and Richard P Wildes. “On diverse asynchronous activity anticipation”. In: *ECCV*. Springer. 2020, pp. 781–799.