J. Integr. Sci. Technol. 2022, 10(1), 11-17 Article

Journal of Integrated
SCIENCE & TECHNOLOGY

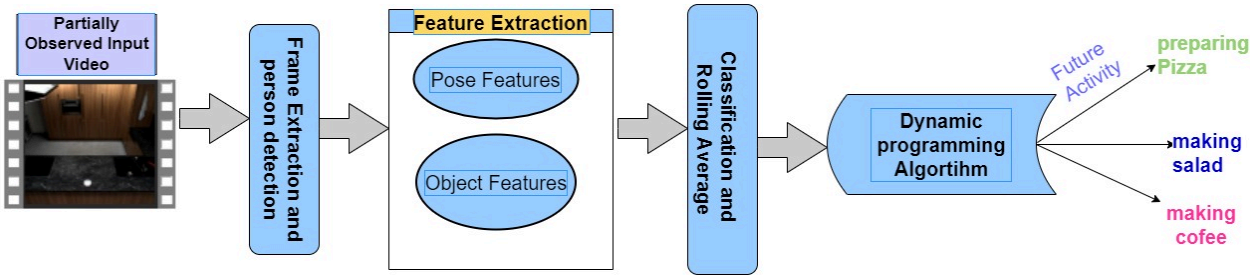
Context aware human activity prediction in videos using Hand-Centric features and Dynamic programming based prediction algorithm

S.N Kakarwal,[1] Ashwini S Gavali[2]

[1]Department of Computer Science and Engineering, PES College of Engineering, Aurangabad, Maharashtra, India, [2]Department of Computer Science and Information Technology, Dr. B.A.Marathwada University, Aurangabad, Maharashtra, India

Received on: 05-Jan-2022, Accepted and published on 24-Feb-2022

ABSTRACT



Activity prediction in videos deals with predicting human activity before it is fully observed. This work presents a context-aware activity prediction approach that can predict long-duration complex human activities from partially observed video. Here, we consider human poses and interacting objects as a context for activity prediction. The major challenges of context-aware activity predictions are to consider different interacting objects and to differentiate visually similar activity classes, such as cutting a tomato and cutting an apple. This article explores the use of hand-centric features for predicting human activity, consisting of various human-object interactions. A Dynamic Programming Based Activity Prediction Algorithm (DPAPA) is proposed for finding the future activity label based on observed actions. The proposed DPAPA algorithm does not employ Markovian dependencies or hierarchical representation of activities and hence is well suited for predicting human activities, which are often non-Markovian and non-hierarchical. We evaluate results on the MPPI Cooking activity dataset, which consists of complex and long-duration activities.

Keywords: context-aware, Activity prediction, Hand-centric features, Interactional object, Dynamic programming

INTRODUCTION

Surveillance cameras are employed everywhere in today's environment to ensure security. Such surveillance systems generate large amounts of video data on a daily basis, but due to a lack of time and human resources, the majority of the data/videos are not adequately analysed. When a criminal case is reported, these recordings are evaluated and analysed by humans, which takes a long time and requires a lot of focus to watch these movies properly. When a criminal incident is discovered in video, the investigative team begins looking for the perpetrator. Arresting a criminal after he has fled the crime scene is time-consuming work.

To address the aforementioned difficulty, several researchers have concentrated on recognizing human activity in videos, and considerable results have been reported in human activity recognition[1]. However, such after-the-fact classification is ineffective in time-critical situations, such as finding the perpetrator after he has fled the crime scene.

The system should predict human intent in advance, allowing for the avoidance of potentially risky behaviour in advance[2]. The goal of this work is to develop a machine vision-based system that can predict and localize suspicious human actions in real-time, as well as leverage previous and current observations to forecast future activity intentions.

A long-term human activity that lasts a long time is made up of a series of actions. This paper refers to action as a specific single movement, while **activity** is a sequence of a variety of **actions**. For example, making a sandwich is an activity that is made up of a series of actions such as cutting tomatoes, carrots, bread, grating butter, and so on. We call this complete activity a **"global activity"**

Corresponding Authors: S.N Kakarwal
Email: s_kakarwal@yahoo.com
Ashwini S Gavali Email: ashwinigavali22@gmail.com

Cite as: *J. Integr. Sci. Technol., 2022, 10(1), 11-17.*

 ISSN: 2321-4635 http://pubs.iscience.in/jist

and each action (sub-activity) **a "local action"**. The goal of this research is to solve the problem of activity prediction by using a series of observed local actions as a trigger to forecast future global actions. We use these local actions as cues for predicting global activity. As the length of the video observation rises, the present local actions are predicted initially, and then the dynamic programming algorithm incrementally determines future possible activity based on past actions and currently seen actions.

The proposed approach keeps track of all past and current local actions performed by individuals while predicting future possible activity, which helps to predict intension more precisely.

This paper extends the You Only Look Once (YOLO)[3] object detection algorithm for detecting and localizing hand-centric features. These features are then used to predict a sequence of observed local actions. A dynamic programming algorithm is proposed for predicting global activity based on observed local action at different progress levels.

LITERATURE SURVEY

In the last few decades, after-the-fact action recognition has been extensively researched, with promising results. State-of-the-art methods[4–9] are capable of precisely labelling the actions, after witnessing the whole action video. However, intelligent systems do not have the luxury of waiting for the complete video in many real-world scenarios (e.g., a vehicle accident or criminal behaviour). For example, system should be able to anticipate a potentially risky driving circumstance rather than realising it after the fact. Unfortunately, most of the existing action recognition approaches are unsuitable for such early prediction tasks.

Activity or action recognition is the task of classifying human activity or actions in a video after they are fully observed, while activity prediction deals with inferring activity before it is fully observed. Such instant-reactive intelligent systems would be very useful in time-sensitive tasks such as criminal spotting, paediatrician action prediction in driverless cars, anticipating dangerous driving situations to avoid accidents, etc.

There are two types of activity predictions: short-term and long-term.

Short-term activities last for a few seconds and do not contain multiple different sub-activities. For example, running, sitting, punching, pushing, handshaking, etc. The prediction task is to simply infer single activities with partially observed frames, e.g., predicting handshaking activity by observing the initial frames of an open arm is easy. This type of prediction can also be called early recognition.

On the other hand, long-term activities last a few minutes and consist of multiple sub-activities. Such approaches aim to predict.

future unobserved activity based on the series of observed local actions. This type of prediction can also be called as intention prediction or future activity prediction[10].

Lots of work has been reported for prediction of shot-term actions using hand crafted features [2,11–14] and Deep Learning based methods[11,15–19.20,21]

Though significant work has been done in early action recognition, very little work has focused on the prediction of long-duration, complex activity prediction. This section presents various available approaches for long-duration activity prediction.

Pei et al.[22] used an And-Or-Graph technique, which incorporates Stochastic Context Sensitive Grammar, for goal inference and intention prediction. They created all feasible parse graphs of a single event by modelling agent-object interactions. The interpretation of the input video is generated by combining all of the possibilities and obtaining the global greatest posterior probability. They also show that by employing hierarchical event contexts, ambiguities in the recognition of atomic actions can be greatly reduced.

Li et al.[23] presented a framework for long-term action prediction by using a Probabilistic Suffix Tree (PST) to capture variable Markov dependencies between action primitives in complex action. Further, they extended their work in [24], to incorporate object context in prediction. Two prediction models have been proposed, namely the action-only model and the context-aware model. The casual relationship between primitive atomic actions is modelled using a Probabilistic Suffix Tree (PST) which can capture small and long order Markov dependencies. Action and object information is encoded as complex symbolic sequences through sequential pattern mining (SPM)

Koppula et al.[25] looked into human action prediction and object affordance. They proposed the ATCRF (anticipatory temporal conditional random field) to explain three types of context information: the hierarchical structure of action primitives, complex spatial-temporal correlations between objects and their affordances, and object and human motion anticipation. ATCRFs are modelled as particles that propagate through time to depict the distribution of probable future actions in order to discover the most likely motion.

Tahmida Mahmud et al.[26] developed a deep neural network for joint prediction of future activity as well as the starting time of the next activity. Object features and activity features were used for representing activity and LSTM network was employed for sequential classification of action. This network can predict the next possible activity based on the last three observed activities.

S Qi et al.[27] argued that human tasks frequently exhibit non-Markovian and compositional properties, and Markov models are insufficient to model such tasks. To deal with this situation, an Earley parser was proposed to parse sequential data in a top-down manner. This generalized Earley parser accepts input from any arbitrary probabilistic classifier and can find the optimal segmentation and labels.

Our goal is to forecast long-duration, complicated, and fine-grained activity using the suggested method before it is completely executed. Similar work was provided in the methodologies for long-term activity prediction cited above. However, these approaches have the following drawbacks:

The work in[23-26] assumes that human actions always have Markov dependencies, although this assumption may not hold true in practise. For example, when creating a salad, the sequence of tasks such as cutting tomatoes, cucumbers, and placing them in a bowl can be completed in any order.

To address the aforementioned problem,[27] portrayed action sequences in a top-down hierarchical manner. However, in the real world, actions may not necessarily follow the same top-down sequence. Take, for example, the following sequence represented by a parser: take a bowl, wash the object, cut the object, and place

them in a bowl. It is possible that a tomato was chopped without being washed. In this situation, the parser will not effectively parse the sequence, and the prediction accuracy may suffer.

Another difference between above cited approaches and proposed approach is that, former approaches considers all present object in scene in prediction while our approach consider only those object which actually grabbed by human with hand. All the above approaches cited above have used motion based features for activity prediction such as dense trajectory features[28], 3d convolutional networks[8], C3D features[5], Inflated 3D ConvNet[29] CNN (I3D) whereas our approach uses appearance based feature extracted using 2d convolutional networks.[3,30]

Our approach is independent of any assumptions made in abovementioned approaches and can incrementally predicts future activity. As the progress level of unobserved activity increases, the prediction results improve over time.

The main objectives of our approach are:

- To predict current human action in video along with bounding box localization.
- Keep track of each local action and predict the label for future activity using the proposed DPAPA.
- Predict fine-grained human actions where human and object interaction is present.

PROPOSED METHODOLOGY

In this section, we discuss the technical details of proposed method. Section i discusses the training procedure of Action detection model and Object detection model. Section ii discuss the Local Action Prediction Model, and section iii discuss the Dynamic Programming Based Activity Prediction Algorithm (DPAPA)

i. Training Phase

In proposed work we trained two deep networks namely Action detection model and Object detection model. The basic steps involve in training phase are depicted in Figure 1. The details of each step are discussed below:

1. ***Keyframe extraction:*** For each action and object class we select representative key frames which represent specific class of action and object precisely.
2. ***Dataset construction****:* To make deep learning model more generalized we added some external data (i.e google images) for each action and object class. Addition of external data significantly improved the performance.
3. ***Image super resolution****:* The scale of the action and object in dataset is very small which is hard to detect and predict. To enhance the quality of image we apply deep learning based image super resolution so that small scale objects would become clearly visible[31].
4. ***Bounding Box Annotation:***

   a)Action representation using hand-centric features: Our approach uses only appearance-based features for predicting local atomic actions. The idea here is that a human pose and interacting objects can together predict any action without the need for motion-based features. Our approach takes advantage of the You Only Look Once (YOLO) algorithm, which is a neural network-based algorithm that is used for real-time object detection. Here, we extended the YOLO algorithm to detect human actions.

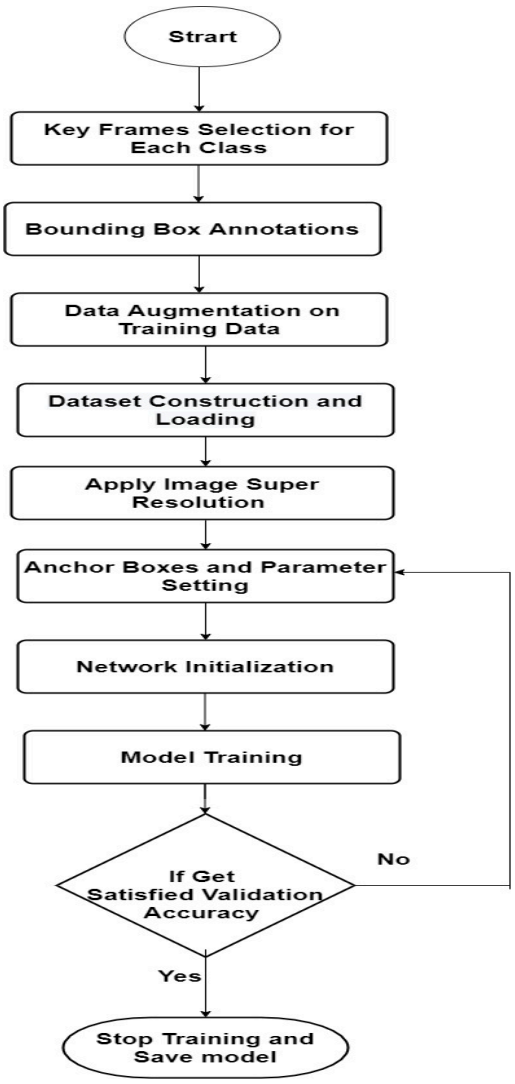


Figure 1: Training flowchart of Action detection and Object detection model

Instead of training the entire human pose, here we focused on hand-centric features for training. The idea behind using hand-centric features is that most daily activities are performed with the help of hands, so training only specific human body parts improved the results on the chosen dataset significantly.

b) Interactional object representation: Instead of considering all objects in the scene, our approach considers only those objects that are grabbed by human hands. The reason behind this is that considering all the objects in the scene may not make sense in predicting as some extra objects that may be present in the scene may degrade generalization performance. But, this approach can also be extended to leg-centric, body-centric features depending on the context.

5. Action detection model training approach: YOLO algorithm with an underlying Resnet-51 architecture is used for training action classes such as cutting, blending, washing etc. Here we used open source implementation of Keras and Tensorflow libraries. The model is trained on the NVIDIA P100 GPU provided by Google Colab. We used image size 900 by 900 as input to model. We used

a stochastic gradient descent with momentum as network optimizer. To prevent overfitting, we used dropout layer with a probability of 0.3 after each layer. We use a batch size of 64 and a learning rate of 0.001. We used data augmentation such as rotation, zooming, width and height shift etc.

6. Object detection model training approach:

As few objects have very less representative frame that resulted While training object models, a few objects have a very representative frame. This results in a class imbalance problem. Class imbalance problems adversely affect model performance. To deal with the class imbalance problem, we used two model approaches as shown in figure 2 for detecting and classifying interactional objects. The first model is YOLO v5, whose responsibility is to just identify the super class (i.e., object) and the second model is an image classifier that takes the desired superclass and further classifies it into subclasses such as tomato, potato, egg etc.

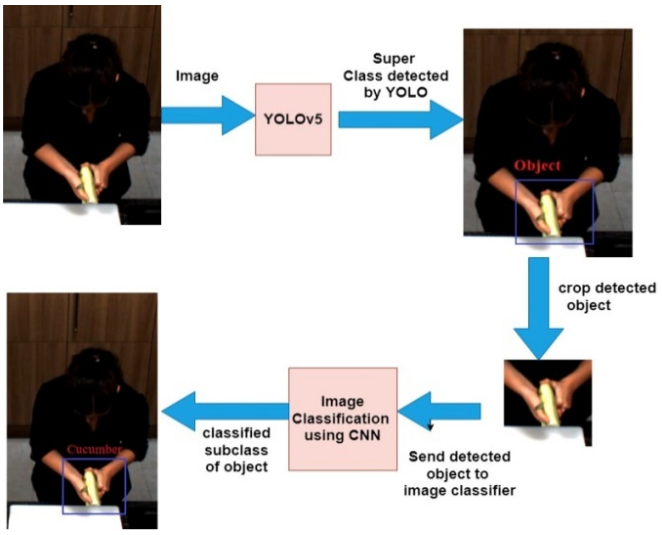


Figure 2: Illustration of two model approach used in Object Prediction

ii. Local Action Prediction Model:

Our Local action prediction model embodies observed action and object labels as cue for future activity prediction. Figure 3 display the overall working of proposed method. Let Σ be the finite set of all possible atomic local actions which are present in training set as

$$\Sigma=\{(a_1,o_1), (a_2,o_2)\ldots\ldots (a_n,o_n)\} \quad (1)$$

Each global activity is represented as sequence of an alphabet of semantic pairs of action and object. Local Action Prediction Model deals with predicting semantic pairs of action and object i.e (a_i, o_i) for each frame f_i.

Algorithm:

Initialization:

Initialize *k*. *k* is the parameter for computing rolling average.

Create $D \in^{IR\ k\times 2}$ represents the matrix with two columns. First column stores the predicted class and second column stores prediction probability.

Create a list *Seq* of length t.

Steps:

Steps 1: Given *O* is the partially observed video up to length t, Extract the frames $[f_1, f_2, f_3 \ldots f_t]$.

Steps 2: For each frame f_i extract Region of Interest (ROI) r_i using pretrained YOLO person detection algorithm [20].

Steps 3: Pass r_i to trained action detection model and get the action label a_i.

Steps 4: Pass r_i to trained object detection model and get the object label o_i.

Steps 5: Combine action and object label as$< a_i , o_i >$ to get local action label al_i. If no object is detected consider action only. Store detected local action label and its probability in *D*.

Steps 6: Compute the Rolling Prediction Average for each predicted class over last k frames as follows:

$$\forall_{j\in 1,2\ldots m}\ Pr(C_j) = 1/n_j \sum_{i=0}^{k} D[i, prob] \quad if\ D[i,class]==C_j \quad (2)$$

Where n_j represents the number of time class C_j appeared in D. Now take the class with highest average probability as follows.

$$Label_i = arg\ max_i\ (Pr) \quad (3)$$

Steps 7: Insert predicted label in *Seq* Store each label only once.

iii. Dynamic Programming Based Activity Prediction Algorithm (DPAPA)

In this section, we present a Dynamic Programming Based Activity Prediction Algorithm (DPAPA) for finding the future activity label that is most likely to occur in future based on the observed sequence of local actions. This algorithm takes a sequence from the base classifier, which is a sequence of actions performed in the partially observed video, as input and predicts possible future activity as output.

Our approach estimates the label for each local activity in observed video and then these local action units are used to predict global class label for ongoing activity. Given the outputs from any probabilistic classifier, the aim is to predict the global activity class C for partially observed video O[1:t],where t is the progress level of partially observed video of length T.

Let *Seq* is list of series of local actions Li predicted from partially observed probe video segments $[O_1, O_2..O_T]$

$$Seq=[L_1, L_2, L_3\ldots\ldots L_t]$$

Let $D_{train} = \{r^1, r^2, r^3, \ldots\ldots r^n\}$ } is the set of training sample set of n different training videos. r_i represents set of all possible local actions for i^{th} training class.

This Prediction score is computed for every Trained Activity class.

Let $Score^{m*2}$ is the matrix of two columns and m rows. Each jth row of matrix stores the posterior probability for class C_j.

Then for each local L_i in the series for observed probe, we computed prediction score for each Training Activity class C_j.as follows:

$$\Pr(C_j) = \sum_{i=1}^{T} P(L_i = Yes \mid Class = C_j) \quad (4)$$

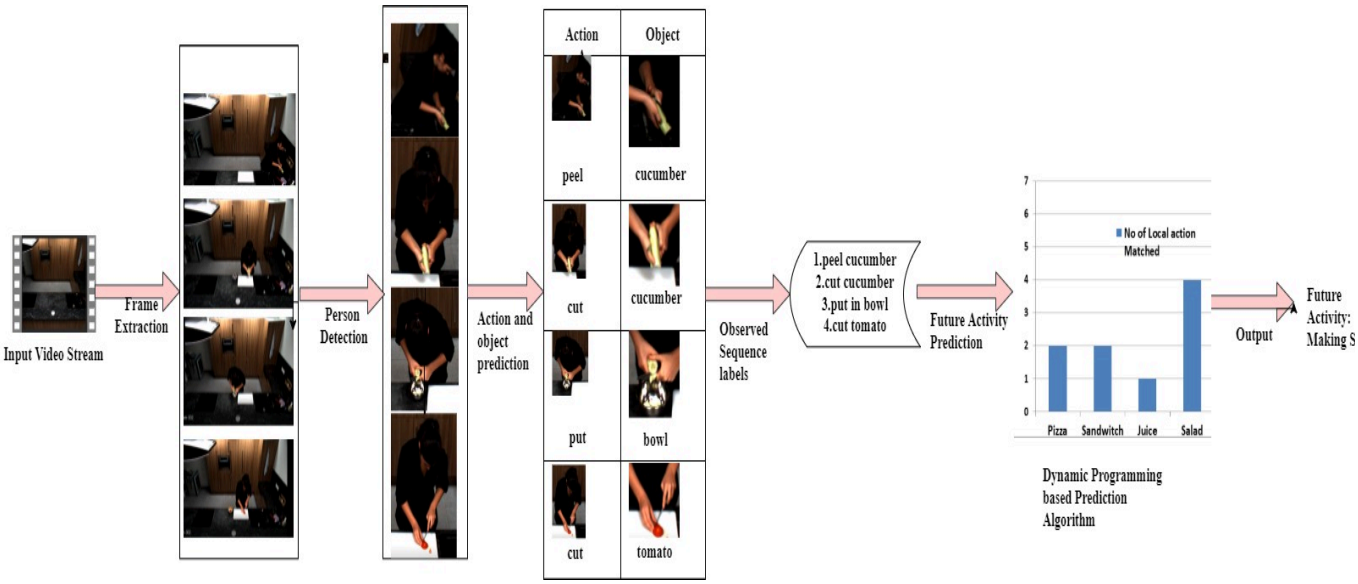


Figure 3: Overall pipeline of proposed work.

This score is the posterior probability that the observed local action with label L_i belong to class C_j Update the probability score for each class as:

$$Score_{j=} Score_{j+} Pr(C_j) \quad (5)$$

This matching process is to be repeated for every action /segment that is detected as video progress. Figure 5 depicts the proposed DPAPA. The idea is take likelihood computed in

previous observation and updates the likelihood for entire observation as new actions are observed. Based on this incremental likehood an optimum activity class is decided that best describes the observed video. Finally the class with maximum Posterior probability is predicted as activity class for ongoing video.

$$FutureLabel= arg\ max_j\ (Score) \quad (6)$$

RESULTS AND DISCUSSION

We evaluated result on MPPI cooking activity dataset using leave-one-out cross validation approach.

MPPI Dataset: The MPPI dataset [32] is a collection of composite activities such as "making a salad," "making a sandwich," and so on. Each composite activity is made up of multiple local actions such as cutting, putting in a pan, stirring, etc.

Participants in each activity class interact with a variety of tools, ingredients, and containers in order to finish a dish. These actions have a longer duration and are complex and finely grained. There are total 65 different local actions classes and total of 5600 video segments are provided for all 65 classes.

There are total 12 subject and total 14 different types of dishes each of which is performed by 3 or 4 subjects. Totally there are 44 videos of length approximately 8 hours. These 44 videos consist of 65 different types of local actions along with 5600 video segment. The task of proposed approach is to predict type of dish subject is preparing before observing complete video.

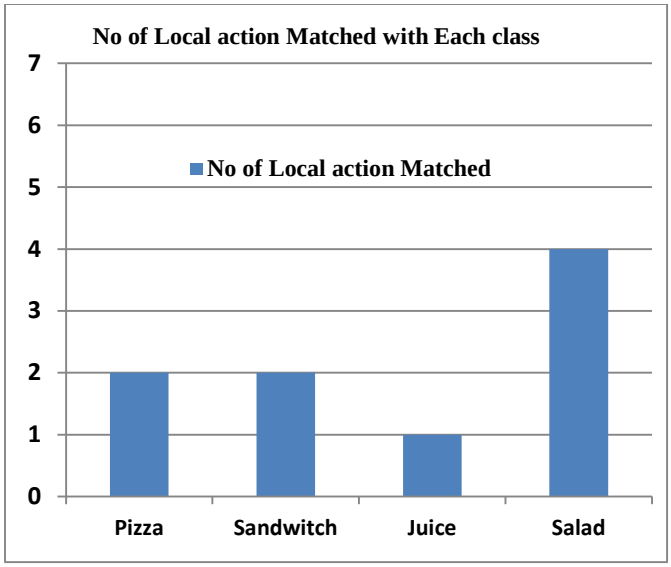
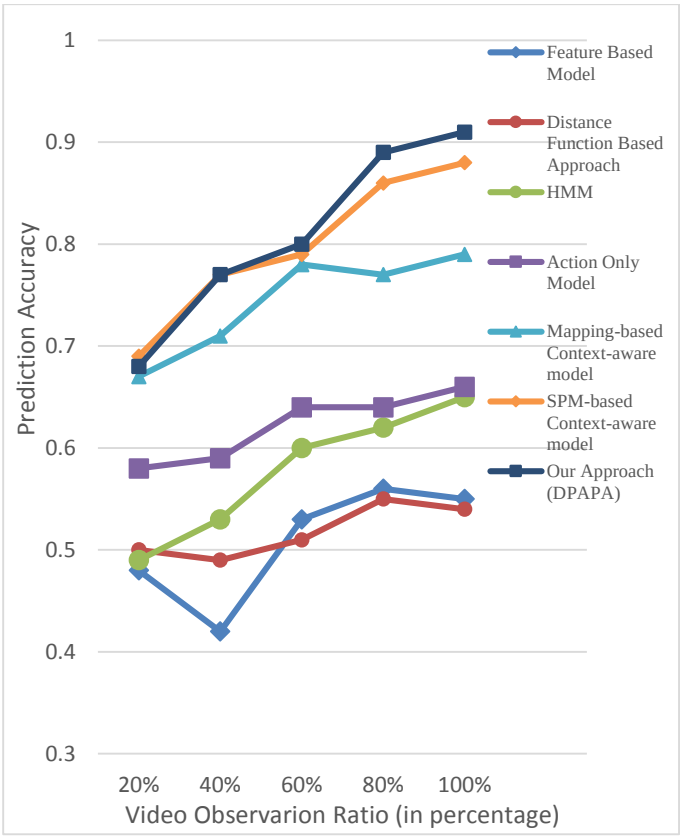


Figure 4: A figure illustrating example of our activity prediction algorithm.

First, in each type of dish, we identified the primary distinguishing local actions. For example, preparing a sandwich involves a set of local actions like washing vegetables, cutting, setting the temperature of the oven, wiping the kitchen, washing hands, etc. However, some of these tasks, such as cleaning and wiping, are common across many dishes. So far, we've just evaluated identifiable action classes when predicting the sort of food. For example, cutting bread, tomatoes, and spreading on bread etc. are the crucial steps in anticipating a sandwich dish.

Results on MPPI Dataset: The table 1 compares the results of our Global Action Prediction model. To show the effectiveness of the proposed method, we compared our results with six other approaches. The comparison is performed at different observation ratio of testing videos. Graph 1 and Table 1 clearly show that our approach performs better as compared to other existing approaches except at observation percentage of 20%.



Graph 1: Performance comparisons of proposed approach.

Table 1: Comparison with State-of-The-Art results on MPPI Dataset.

Sr.No	Methods	Observation Ratio of Videos				
		20%	40%	60%	80%	100%
1	Feature Based Model [24]	0.48	0.42	0.53	0.56	0.55
2	Distance Function Based Approach [24]	0.5	0.49	0.51	0.55	0.54
3	HMM [24]	0.49	0.53	0.60	0.62	0.65
4	Action Only Model [24]	0.58	0.59	0.64	0.64	0.66
5	Mapping-based Context-aware model [24]	0.67	0.71	0.78	0.77	0.79
6	SPM-based Context-aware model [24]	0.69	0.77	0.79	0.86	0.88
7	**Our Approach (DPAPA)**	**0.68**	**0.77**	**0.80**	**0.89**	**0.91**

Conclusion and Future Scope

In this paper, we proposed a novel approach to predict long duration, complex and fine-grained activity. The major contribution includes the use of hand-centric features and interactional objects for modelling activities. A Dynamic programming based Activity Prediction Algorithm (DPAPA) is developed, which makes our approach suitable for predicting activities that are non-Markovian and non-hierarchical. Our approach does not rely on any temporal decomposition, and it can also tolerate noisy local actions. This approach is well suited for predicting activity in real time as it can process 25–30 frames per second.

In future, we will extend this model to track and predict multiple person activities in video.

Conflict of Interest

Authors declared no conflict of interest exist.

References

1. P. Pareek, A. Thakkar. A survey on video-based Human Action Recognition: recent updates, datasets, challenges, and applications; Springer Netherlands, **2021**; Vol. 54.
2. M.S. Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. *Proc. IEEE Int. Conf. Comput. Vis.* **2011**, No. Iccv, 1036–1043.
3. J. Redmon, A. Farhadi. YOLO9000: Better, faster, stronger. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*; Institute of Electrical and Electronics Engineers Inc., **2017**; Vol. 2017-January, pp 6517–6525.
4. D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri. Learning spatiotemporal features with 3D convolutional networks. *Proc. IEEE Int. Conf. Comput. Vis.* **2015**, 2015 Inter, 4489–4497.
5. H. Kataoka, T. Wakamiya, K. Hara, Y. Satoh. Would Mega-scale Datasets Further Enhance Spatiotemporal 3D CNNs? **2020**.
6. Q. Li, H. Cheng, Y. Zhou, G. Huo. Human Action Recognition Using Improved Salient Dense Trajectories. *Comput. Intell. Neurosci.* **2016**, 2016.
7. J. Donahue, L.A. Hendricks, S. Guadarrama, et al. Long-term recurrent convolutional networks for visual recognition and description. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* **2015**, 07-12-June, 2625–2634.
8. S. Ji, W. Xu, M. Yang, K. Yu. 3D Convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, 35 (1), 221–231.
9. K. Simonyan, A. Zisserman. Two-stream convolutional networks for action recognition in videos. *Adv. Neural Inf. Process. Syst.* **2014**, 1 (January), 568–576.
10. Y. Kong, Y. Fu. Human Action Recognition and Prediction: A Survey. **2018**, 13 (9).
11. T. Lan, T.C. Chen, S. Savarese. A hierarchical representation for future action prediction. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **2014**, 8691 LNCS (PART 3), 689–704.
12. M. Hoai, F. De La Torre. Max-margin early event detectors. *Int. J. Comput. Vis.* **2014**, 107 (2), 191–202.
13. Y. Cao, D. Barrett, A. Barbu, et al. Recognize human activities from partially observed videos. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* **2013**, 1, 2658–2665.
14. Y. Kong, Y. Fu. Max-margin action prediction machine. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, 38 (9), 1844–1858.
15. S. Ma, L. Sigal, S. Sclaroff. Learning Activity Progression in LSTMs for Activity Detection and Early Detection. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* **2016**, 2016-December, 1942–1950.
16. Y. Kong, Z. Tao, Y. Fu. Adversarial Action Prediction Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, 42 (3), 539–553.

17. X. Wang, J.F. Hu, J.H. Lai, J. Zhang, W.S. Zheng. Progressive teacher-student learning for early action prediction. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* **2019**, 2019-June, 3551–3560.
18. K. Soomro, H. Idrees, M. Shah. Online Localization and Prediction of Actions and Interactions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, 41 (2), 459–472.
19. G. Singh, S. Saha, M. Sapienza, P. Torr, F. Cuzzolin. Online Real-Time Multiple Spatiotemporal Action Localisation and Prediction. *Proc. IEEE Int. Conf. Comput. Vis.* **2017**, 2017-Octob, 3657–3666.
20. H. Gammulle, S. Denman, S. Sridharan, C. Fookes. Predicting the future: A jointly learnt model for action anticipation. *Proc. IEEE Int. Conf. Comput. Vis.* **2019**, 2019-Octob, 5561–5570.
21. P. Lei, S. Todorovic. Temporal Deformable Residual Networks for Action Segmentation in Videos. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* **2018**, 6742–6751.
22. M. Pei, Z. Si, B.Z. Yao, S.C. Zhu. Learning and parsing video events with goal and intent prediction. *Comput. Vis. Image Underst.* **2013**, 117 (10), 1369–1383.
23. K. Li, J. Hu, Y. Fu. Modeling complex temporal composition of actionlets for activity prediction. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **2012**, 7572 LNCS (PART 1), 286–299.
24. K. Li, Y. Fu. Prediction of human activity by discovering temporal sequence patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, 36 (8), 1644–1657.
25. H.S. Koppula, A. Saxena. Anticipating Human Activities Using Object Affordances for Reactive Robotic Response. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, 38 (1), 14–29.
26. T. Mahmud, M. Billah, M. Hasan, A.K. Roy-Chowdhury. Prediction and Description of Near-Future Activities in Video. *Comput. Vis. Image Underst.* **2021**, 210.
27. S. Qi, B. Jia, S. Huang, P. Wei, S.C. Zhu. A Generalized Earley Parser for Human Activity Parsing and Prediction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, 43 (8), 2538–2554.
28. H. Wang, C. Schmid. Action recognition with improved trajectories. *Proc. IEEE Int. Conf. Comput. Vis.* **2013**, 3551–3558.
29. J. Carreira, A. Zisserman. Quo Vadis, action recognition? A new model and the kinetics dataset. *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017* **2017**, 2017-January, 4724–4733.
30. K. He, X. Zhang, S. Ren, J. Sun. Deep residual learning for image recognition. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* **2016**, 2016-Decem, 770–778.
31. C. Dong, C.C. Loy, X. Tang. Accelerating the super-resolution convolutional neural network. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **2016**, 9906 LNCS, 391–407.
32. M. Rohrbach, S. Amin, M. Andriluka, B. Schiele. A database for fine grained activity detection of cooking activities. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* **2012**, 1194–1201.