# Towards Action Prediction Applying Deep Learning

1rt Marjorie Chalen T.
*Escuela Superior Politécnica del Litoral*
Guayaquil, Ecuador
mchalen@espol.edu.ec

2nd Boris Vintimilla
*Escuela Superior Politécnica del Litoral*
Guayaquil, Ecuador
bvintimi@espol.edu.ec

***Abstract*—Considering the incremental development future action prediction by video analysis task of computer vision where it is done based upon incomplete action executions. Deep learning is playing an important role in this task framework. Thus, this paper describes recently techniques and pertinent datasets utilized in human action prediction task.**

***Index Terms*—action prediction, early recognition, early detection, action anticipation, cnn, deep learning, rnn, lstm.**

## I. Introduction

The Human Action Recognition, HAR, and Future Action Prediction, FAP, from videos has become an important topic in computer vision due to its broad scope of applications. Its most relevant applications are video surveillance, human-machine interaction, video analysis in sports and health care. Detection of suspicious or abnormal activities in automated video surveillance systems is required in places such as airports and subway stations. In areas of health care of a patient, child or elderly person, it allows real time monitoring.

Computer vision contemplates the recognition of an action as the identification of a defined pattern and then assign it a label defining the type of action. There are four different ways to categorize an action: gestures, actions, interactions and group activities [1]. Gestures are movements related to a specific part of the body and is considered an atomic action. Action is a type of activity that is performed by a single person. Interaction is a type of activity carried out by two actors, where an actor can be a person or object. Group Activities are more complex, consist of a combination of gestures, actions and interactions, with more than two humans and one or more objects.

In action prediction, the action label is inferred from incomplete observations of the action [2]. There are three categories in action prediction: Early action recognition, Early action detection and Action anticipation that Ke [3] defined as the following way:

- Early action recognition recognizes the label of an action from a partial observation of this action.
- Early action detection aims to detect an action as early as possible before the action ends from untrimmed videos.
- Action anticipation corresponds to anticipation of immediate future after the observation.

Action anticipation also could be named directly as prediction where a label of an action/activity before it is fully performed.

The objective of this paper is to bring a simple understanding of the techniques and dataset utilized in future action prediction.

The sections below describe the execution of this project: II – Background description and related work III – Data analysis methods and techniques IV – Datasets evolution, and V – Current status of the research and future works.

## II. Background

### A. Related Work

Human action prediction have been presented such as [4] bringing a generic feature and classification approach. Classifying video representation as Low-Level, Mid-Level and Unsupervised Features. The first, are visual information on both spatial and temporal domain. The second, are usually built from the first and have some semantic meaning. The third, uses deep learning to extract spatio-temporal features. Then, establishes a model classification similar to activity recognition approach [5] where it consider discriminative model, generative model and deep neural model. It also presents more relevant datasets for this topic of research. This paper pretends to make a refresh of those classifications and a dataset review focused on last paper works presented in relevant conference.

Another work [6] focus its attention in Human-Human Interactions where some challenges are identified such as variation in visual appearance where people interactions could be observed in many different environments and conditions (illumination condition, dynamic background), intra-class variation in interaction performance meaning people behave differently for the same actions, and data collection challenges and labeling, there is no common labeling of the interaction classes.

### B. Other Challenges

*a) Inter-class Variation:* there are similarities in different action categories. For instance, walking and running have similar patterns of motion could incurre in misclassification.

*b) Camera motion:* could affect feature extraction, it should be modeled and compensated. Although, viewpoints changes affects feature extractions.

Additionally, One of the main problems in action prediction are lot of redundant frames, and discriminating frames that may appear anywhere in the video.

## III. Methods and Techniques

Some approaches have been proposed obtaining acceptable results. A typical approach in human activity recognition is a spatial-temporal representation considering different types of information such as person and objects appearance features, contextual importance, interactions with objects, and human pose. However, these methods are not enough for prediction, some recently works have incorporated human behavioral information, and interaction with their surroundings to get this goal.

It is important to mention that since using Deep Neural Network in predicting future activities there have been different features approaches such as create a new type of image as an input of the learning model, combining deep features with handcrafted features.

Future action prediction is similar to action recognition [7] since its task could be simplified by two stages: Action representation and Action classification where the first is the feature representation and extraction. The second, is the early action recognition, early action detection or action anticipation (prediction).

### A. Feature extraction

*a) Convolutional Neural Network, CNN:* Most works using generic feature extracted with pre-trained CNN from ImageNet have been presented.

- Deep temporal features [8] [9] [10] [11], flow coding images are computed from consecutive video frames.
- Context-Aware Feature Extraction [12]. This sub-model is similar to VGG-16 from conv5-3 up to the last fully connected layer, with the number of units in the last fully connected layer changed from 1000 to the number of activities N. In essence, this sub-model focuses on extracting a deep representation of the whole scene for each activity and thus incorporates context.
- OpenPose, local patches of skeleton proposals which are decided by the joints of the skeleton. At every joint, spatial feature were extracted with pre-trained model of VGG-16 [13].
- AlexNet CNN (pre-trained on ImageNet) architecture is adopted to extract visual features [14].
- Dynamic CNN from dynamic images created from observed RGB data where dynamic images summarise temporal evolution of appearance of few frames into a single image, so it captures motion information of those frames [15].
- Person behavior module extracts visual information from the behavioral sequence of the person. Appearance (ROIAlign - Mask R-CNN) and body movement (keypoint linear transformation). Person interaction module looks at the interaction between a person and their surroundings. Person-scene (CNN) and person-object (geometric relation and object type, one-hot encode, model) [16].
- SSNet based on ConvNet over tree structure [17]. Tree convolutions in spatial domain to learn the multi-level (local, mid-level, and holistic) structured representations.

*b) CNN and Handcrafted:* Combination of deep extracted features and handcrafted features.

- Spatiotemporal interest points (STIPs) [18] with 3DCN and dense trajectory features (DTs) [19]. Bag-of-words model is used to encode is adopted to encode STIPs and DTs features.

### B. Recurrent Neural Network, RNN

*a) Early Activity Recognition:*

- RNN model [20] that performs a structured prediction over the label hierarchy (structRNN).
- Multiple Soft labels Recurrent Neural Network, MSRNN, where depth patches around each body part and relative skeleton features for each frame [21].

*b) Action Anticipation / Prediction:*

- Dynamic CNN [15].
- Multi-Sacle temporal convolution [22].

### C. Long Short Term Memory, LSTM

*a) Early Activity Detection:*

- LSTM to process the sequence and learn the structural models from global and local interaction contexts in a sequential order [8].

*b) Action Anticipation / Prediction:*

- LSTM to incorporate sequential activity context [23].
- Encoder-decoder network that uses a LSTM network as basic cell, with a loss function as activity classification based on probability. score [24] [16].

### D. Linear Regression

*a) Action Anticipation / Prediction:*

- Multiple Regression Network, Linear regression on unlabeled repository to predict fc7 (fully connected seven layer) in the future [14].

## IV. Datasets Evolution

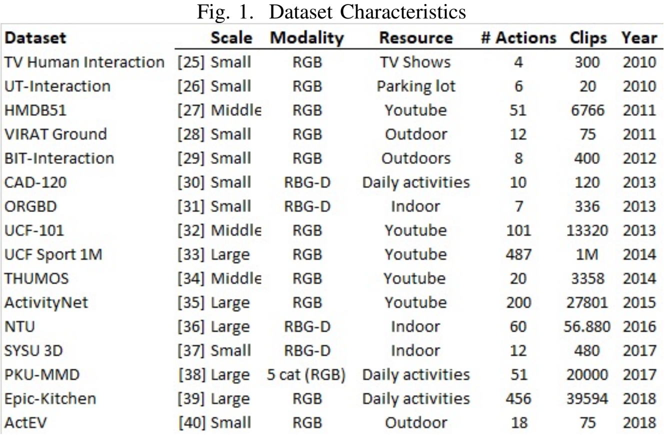
Datasets used in human activity prediction researchers have incremented not only number of actions classification, also number of video clips containing those actions ascending to thousands of video instances, making those datasets more challenged.

Table 1 is presenting more relevant recently used datasets. Other than RBG dataset categories are being developed such as skeleton annotations, egocentric viewpoint/perspective (head mounted camera in Epic-Kitchen).

## V. Conclusions and Future Work

This research area is still in development with many promising results. A benchmark performance analysis of the reviewed algorithms with recently available datasets is needed. Also, more algorithms that pretend to achieve the best accurate and early predictions for most of all actions will be developed.

Fig. 1. Dataset Characteristics



| Dataset | Scale | Modality | Resource | # Actions | Clips | Year |
|---|---|---|---|---|---|---|
| TV Human Interaction | [25] Small | RGB | TV Shows | 4 | 300 | 2010 |
| UT-Interaction | [26] Small | RGB | Parking lot | 6 | 20 | 2010 |
| HMDB51 | [27] Middle | RGB | Youtube | 51 | 6766 | 2011 |
| VIRAT Ground | [28] Small | RGB | Outdoor | 12 | 75 | 2011 |
| BIT-Interaction | [29] Small | RGB | Outdoors | 8 | 400 | 2012 |
| CAD-120 | [30] Small | RBG-D | Daily activities | 10 | 120 | 2013 |
| ORGBD | [31] Small | RBG-D | Indoor | 7 | 336 | 2013 |
| UCF-101 | [32] Middle | RGB | Youtube | 101 | 13320 | 2013 |
| UCF Sport 1M | [33] Large | RGB | Youtube | 487 | 1M | 2014 |
| THUMOS | [34] Middle | RGB | Youtube | 20 | 3358 | 2014 |
| ActivityNet | [35] Large | RGB | Youtube | 200 | 27801 | 2015 |
| NTU | [36] Large | RBG-D | Indoor | 60 | 56.880 | 2016 |
| SYSU 3D | [37] Small | RBG-D | Indoor | 12 | 480 | 2017 |
| PKU-MMD | [38] Large | 5 cat (RGB) | Daily activities | 51 | 20000 | 2017 |
| Epic-Kitchen | [39] Large | RGB | Daily activities | 456 | 39594 | 2018 |
| ActEV | [40] Small | RGB | Outdoor | 18 | 75 | 2018 |

## REFERENCES

[1] J. K. Aggarwal and M. S. Ryoo, "Human Activity Analysis: A Review," ACM Comput. Surv., vol. 43, no. 3, p. Article No. 16, 2011.

[2] M. Ryoo. "Human activity prediction: Early recognition of ongoing activities from streaming videos," ICCV, 2011.

[3] Q.Ke, M. Fritz and B. Schiele "Time-Conditioned Action Anticipation in One Shot," CVPR, 2019.

[4] S. Zhang, Z. Wei, J. Nie, L. Huang, S. Wang, and Z. Li, "A Review on Human Activity Recognition Using Vision-Based Method," J. Healthc. Eng., vol. 2017, 2017.

[5] ...

[6] A. Stergiou. and R. Poppe. "Analyzing Human-Human Interactions: A Survey,". Computer Vision and Image Understanding, Volume 188, November 2019, 102799.

[7] J. K. Aggarwal and M. S. Ryoo, "Human Activity Analysis: A Review" ACM Comput. Surv., vol. 43, no. 3, p. Article No. 16, 2011.

[8] Y. Kong, S. Gao, B. Sun, and Y. Fu. "Action prediction from videos via memorizing hard-to-predict samples," AAAI, 2018.

[9] A. Bux, "Vision-based Human Action Recognition using Machine Learning Techniques by Allah Bux School of Computing and Communications December 2017 Declaration of Authorship," no. December, 2017.

[10] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid. "Leveraging structural context models and ranking score fusion for human interaction prediction," IEEE Transactions on Multimedia, 20(7):1712–1723, 2018.

[11] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. "Temporal segment networks: Towards good practices for deep action recognition," ECCV, 2016.

[12] Q. Ke, M. Bennamoun, S. An, F. Boussaid, and F. Sohel. "Human interaction prediction using deep temporal features," ECCV, 2016.

[13] L. Chen, J. Lu, Z. Song, and J. Zhou. "Part-activated deep reinforcement learning for action prediction," ECCV, 2018.

[14] C. Vondrick, H. Pirsiavash, and A. Torralba.. "Anticipating visual representations from unlabeled video.," CVPR, 2016.

[15] C. Rodriguez, B. Fernando. "Action Anticipation By Predicting Future Dynamic Image," ECCV, 2018.

[16] J. Liang, Lu Jiang, J. Niebles, A. Hauptmann, L. Fei-Fei. "Peeking into the Future: Predicting Future Person Activities and Locations in Videos," CVPR, 2019.

[17] J. Liu, A. Shahroudy, G. Wang, L.-Y. Duan, and A. C. Kot. "Skeleton-based online action prediction using scale selection network," TPAMI, 2019.

[18] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. "Behavior recognition via sparse spatio-temporal features," In VS-PETS, 2005.

[19] H.Wang, A. Kl¨aser, C. Schmid, and C.-L. Liu. "Action recognition by dense trajectories," In CVPR, pages 3169–3176, 2011.

[20] W. Li and M. Fritz. "Recognition of ongoing complex activities by sequence prediction over a hierarchical label space," CACV, 2016.

[21] J.-F. Hu, W.-S. Zheng, L. Ma, G. Wang, J.-H. Lai, and J. Zhang. "Early action prediction by soft regression," TPAMI, 2018

[22] Q.Ke, M. Fritz and B. Schiele. "Time-Conditioned Action Anticipation in One Shot," CVPR, 2019.

[23] T. Mahmud, M. Hasan, and A. K. Roy-Chowdhury. "Joint prediction of activity labels and starting times in untrimmed videos,"ICCV, 2017.

[24] J. Gao, Z. Yang, and R. Nevatia. "Red: Reinforced encoder decoder networks for action anticipation," BMPR, 2017.

[25] Ryoo, M.S. "Human activity prediction: Early recognition of ongoing activities from streaming videos," ICCV, 2011.

[26] Ryoo, M.S., Aggarwal, J.K. " UT-Interaction Dataset," ICPR contest on Semantic Description of Human Activities (SDHA), 2010. http://cvrc.ece.utexas.edu/ SDHA2010/Human Interaction.html

[27] Kuehne, H., Jhuang, H., Garrote, E., Poggio, T.A., Serre, T. "HMDB: A large video database for human motion recognition," in ICCV, 2011. 2556–2563.

[28] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. Aggarwal, H. Lee, L. Davis, et al. A large-scale benchmark dataset for event recognition in surveillance video. In CVPR, 2011.

[29] Kong, Y., Jia, Y., Fu, Y. "Learning human interaction by interactive phrases," ECCV. pp. 300–313, 2012.

[30] Koppula, H.S., Gupta, R. Saxena, A. "Learning human activities and object affordances from RGB-D videos," CoRR, abs/1210.1207, 2012)

[31] S. Koppula, R. Gupta, A. Saxena. "Learning Human Activities and Object Affordances from RGB-D Videos," International Journal of Robotics Research (IJRR), in press, Jan 2013.

[32] 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.

[33] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. "Large-scale video classification with convolutional neural networks," in CVPR, 2014.

[34] Jiang, Y.G., Liu, J., Zamir, A.R., Toderici, G., Laptev, I., Shah, M., Sukthankar, R. "THUMOS challenge: Action recognition with a large number of classes," http://crcv.ucf.edu/THUMOS14/, 2014.

[35] Activitynet challenge 2016. http://activity-net.org/challenges/2016/, 2016.

[36] A. Shahroudy, J.Liu, T. Ng, G. Wang. "NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis," arXiv, 2016.

[37] Jian-Fang Hu, Wei-Shi Zheng, Jianhuang Lai, Jianguo Zhang, "Jointly Learning Heterogeneous Features for RGB-D Activity Recognition," IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol.39 (11), 2017, pp.2186-2200.

[38] C. Liu, Y. Hu, Y. Li, S. Song, and J. Liu. "Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding," arXiv, 2017.

[39] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, et al. "Scaling egocentric vision: The epic-kitchens dataset," In European Conference on Computer Vision, 2018.

[40] G. Awad, A. Butt, K. Curtis, J. Fiscus, A. Godil, A. F. Smeaton, Y. Graham, W. Kraaij, G. Qunot, J. Magalhaes, D. Semedo, and S. Blasi. Trecvid 2018: Benchmarking video activity detection, video captioning and matching, video storytelling linking and video search. In TRECVID, 2018.