End of Semester Progress Report:
Future Activity and Duration Prediction in Real-Time

Aazia Azmi
aaziaazmi@gatech.edu

Abstract:

Recent advances in Human Activity Recognition have achieved segmenting and classifying video frames with high levels of accuracy in just a few seconds. Predicting long-term activities is a less explored problem space and my work this semester aims to explore this field through different datasets and models that can make predictions on what activities a person will perform next and what their durations will be.

Introduction:

Last semester, I learned how Machine Learning can be used to predict what action a person is performing in real-time by making predictions based on multi-modal data using LSTMs [1]. Such models can have many practical applications and can offer insights to machines that are observing activities in their environment. My aim this semester, is to take these models a step further and investigate methods to predict not only the activity a person is performing, but also what the rest of their tasks will be and how long they will spend on them.

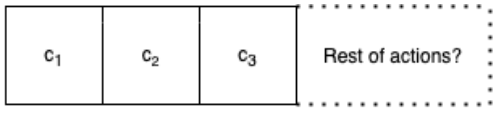

Figure 1: What I want to predict

[2] is currently one of the very few publications that explore this field. To answer the question “What will you do when?”, using [2] as the main reference, I explored CNN and RNN models trained on the 50Salads [6], Breakfast [7], and JIGSAWS [8] datasets. Because [3] agrees that encoding input data and then decoding the output achieves better results than models trained directly using RGB data, I have encoded all my input data before training.

Based on some observations on the JIGSAWS dataset that were made by implementing unsupervised learning algorithms, I found a correlation between workers’ skill level and average activity duration. I explored how changing the encoding of input data to accommodate this feature can improve the model’s accuracy.

Next, I explored how skill level or action quality can be identified in addition to activity labels in real-time and reviewed some real-time classifiers like RNN-HMMs, MS-TCNs, and LSTMs that could be suitable for the same.

Related Work:

Re-implementing Abu Farha’s “Anticipating Temporal Occurrences of Activities” paper [2] gives us both a CNN and an RNN approach to solving this problem. Both use a two-step approach. For RNN, first, we encode the frames into tuples that contain the length of the segment observed and its label as 1-hot encoding. Then we implement our future-predictor on these tuples. The models trained on such tuples came out as 5% more accurate than models trained directly on observed frames. For this study, we ignore the Activity Recognition model seen in Figure 2 and work directly with ground truth values for training our models.

The RNN approach uses the one hot class-encoding tuples discussed earlier as input and tuples with the remaining time of the last observed action and the label and duration of the next

action to follow as output. The output is recursively combined with the input to predict all future actions. The model has 2 stacked layers of 256 gated recurrent units and fully connected layers at the input and output. For the output layer for both length predictions, a rectified linear unit is used to ensure positive length outputs.

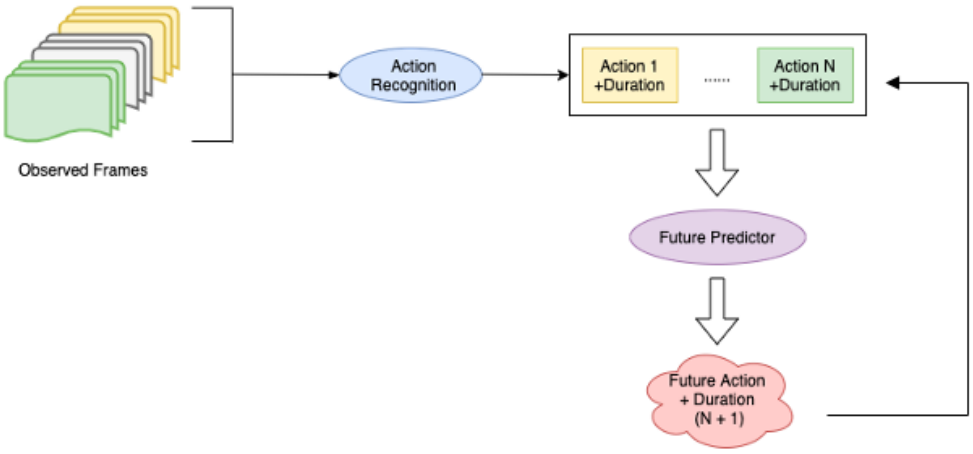

Figure 2: Sample System for RNN-based Future predictor

The CNN approach aims to predict all actions directly in one single step. The input is encoded into a matrix where columns represent action classes and rows represent action segments. The CNN consists of two convolutional layers and two fully connected layers. A 1-D Gaussian filter is applied along each column of the output for temporal smoothing.

Unfortunately, if the RNN outputs an erroneous prediction at some point, this error is likely to propagate through time. The CNN, on the contrary, uses the observed part of the video to predict all future actions directly, so errors are less likely to propagate from one segment to another. As observed in Table 1, the RNN model outperforms the CNN model for shorter duration predictions, but the CNN model performs equally well or better for longer duration predictions.

My Approach:

[2] was implemented on the Breakfast and 50Salads datasets only. To incorporate another feature into our tuples – activity quality, I explored the JIGSAWS dataset.

In the JIGSAWS dataset, analyzing how much time is spent by workers of different skill levels (Expert, Intermediate, Novice) on different actions, we can see a direct correlation between skill level and activity duration as depicted in Figure 3. In particular, it is observed that high skill workers spend more time on their actions than lower-skilled workers performing the same action.

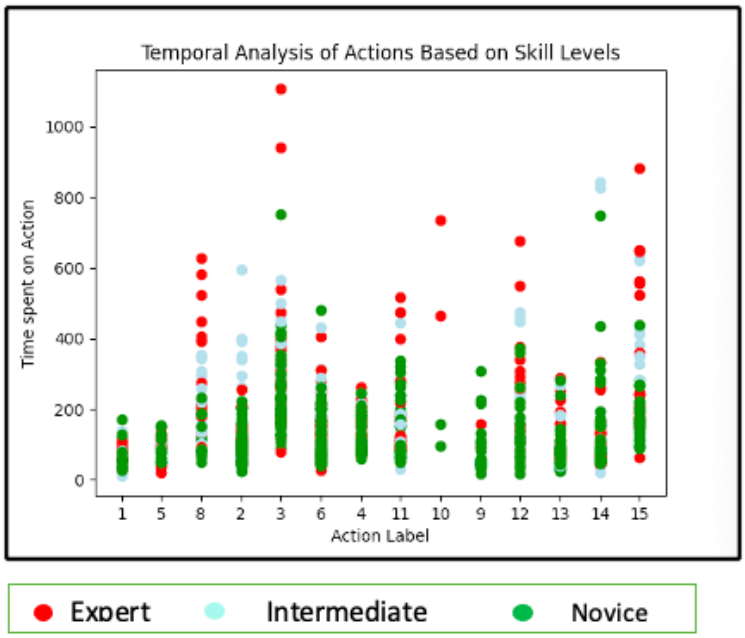

Figure 3: Action vs Duration graph for different skill level workers

Based on this analysis, a slight change is made to our existing RNN model to include the skill feature.

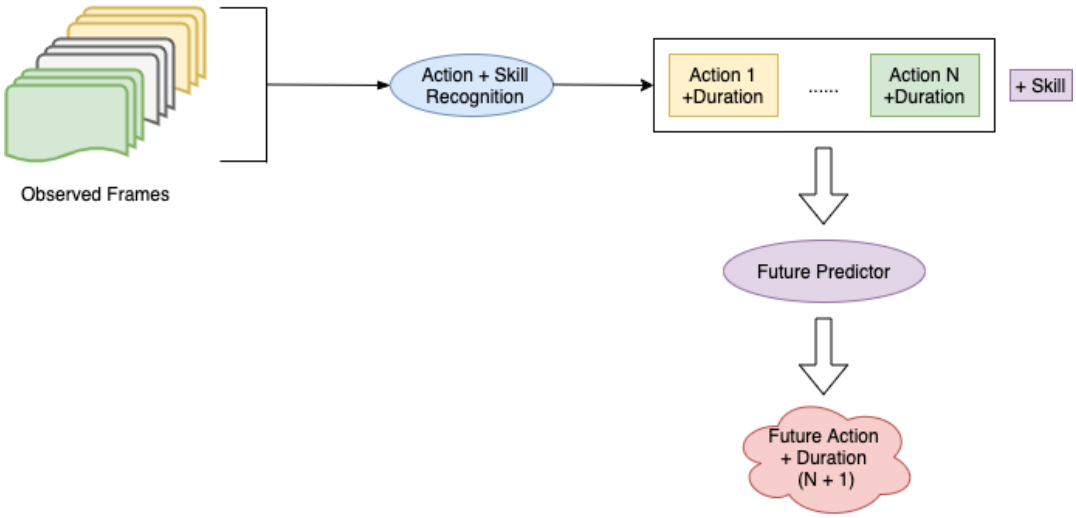

Figure 4: New sample system for RNN-based Future Predictor

To accommodate this model, our action recognition model must also be able to make real-time predictions about skill level. However, we will still simply use ground-truth values for training thus ignoring the Activity + Skill Recognition in Figure 4. The action segments in the JIGSAWS dataset are of slightly longer duration and are more suitable for a CNN model, for easier implementation reasons, I applied an RNN model on this dataset. In this new model, skill is added into the encoded input tuples.

As seen in Table 2, incorporating the skill feature significantly increases the model's accuracy by an average of 0.0261. The increase in accuracy is more significant in shorter duration predictions than longer ones.

Further studies

As mentioned before, for easier study, only ground-truth values are used for training. Practically, there would be an activity (and skill) recognition model that would observe activity in real-time. Farha uses an RNN-HMM hybrid model resembling [4] for this purpose. I

implemented an MS-TCN [5] model and it yielded an 80% accuracy on the 50Salads dataset for normal Human Activity Recognition. I am yet to implement a model that also predict skill levels. Because the output from the activity (and skill) recognition models cannot be perfect, we can expect to see a reduction of accuracies in practical application.

It's also important to figure out how the added skill feature can be accommodated in our existing CNN model and how different data processing methods can improve our models.

Conclusion:

This semester I had the chance to experiment with existing future prediction models and see the result of my experimentation, ie: including skill feature enhances the model's accuracy. I also had the chance to start exploring real-time skill prediction methods – something I hope to pick up on next semester.

Evaluation:

Observation %	20%				30%			
Prediction %	10%	20%	30%	50%	10%	20%	30%	50%
Breakfast RNN	0.6035	0.5044	0.4528	0.4042	0.6145	0.5025	0.4490	0.4175
Breakfast CNN	0.5797	0.4912	0.4403	0.3926	0.6032	0.5014	0.4518	0.4051
50Salads RNN	0.4320	0.3119	0.2552	0.1682	0.4419	0.2951	0.1996	0.1038
50Salads CNN	0.3608	0.2762	0.2143	0.1548	0.3736	0.2478	0.2078	0.1405

Table 1: Observations recorded when the RNN and CNN models are trained on the first 20% and 30% of the datasets and make predictions on the next 10%, 20%, 30%, and 50% of data

Observation %	20%				30%			
Prediction %	10%	20%	30%	50%	10%	20%	30%	50%
Old RNN model	0.5727	0.4895	0.4324	0.4042	0.5932	0.5210	0.4473	0.4112
New RNN model	0.6334	0.5117	0.4615	0.4010	0.6371	0.5332	0.4865	0.4190

Table 2: Observations recorded when we train our model on the first 20% and 30% of the JIGSAWS dataset and make predictions on the next 10%, 20%, 30%, and 50% of data.

References:

1. Francisco Javier Ordonez, Daniel Roggen. Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. In MDPI, 2016.
2. Yazan Abu Farha, Alexander Richard, and Juergen Gall. When will you do what? - Anticipating Temporal Occurrences of Activities. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
3. M. Garbade and J. Gall. Thinking outside the box: Spatial anticipation of semantic categories. In British Machine Vision Conference (BMVC), 2017.
4. A. Richard, H. Kuehne, and J. Gall. Weakly supervised action learning with RNN based fine-to-coarse modeling. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
5. Yazan Abu Farha and Juergen Gall. MS-TCN: Multi-Stage Temporal Convolutional Network for Action Segmentation

6. S. Stein and S. J. McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In ACM International Joint Conference on Pervasive and Ubiquitous Computing, 2013.
7. H. Kuehne, A. Arslan, and T. Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 780–787, 2014.
8. Yixin Gao, S. Swaroop Vedula, Carol E. Reiley, Narges Ahmidi, Balakrishnan Varadarajan, Henry C. Lin, Lingling Tao, Luca Zappella, Benjamın Bejar, David D. Yuh, Chi Chiung Grace Chen, Ren´e Vidal, Sanjeev Khudanpur and Gregory D. Hager, The JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS): A Surgical Activity Dataset for Human Motion Modeling, In Modeling and Monitoring of Computer Assisted Interventions (M2CAI) – MICCAI Workshop, 2014.