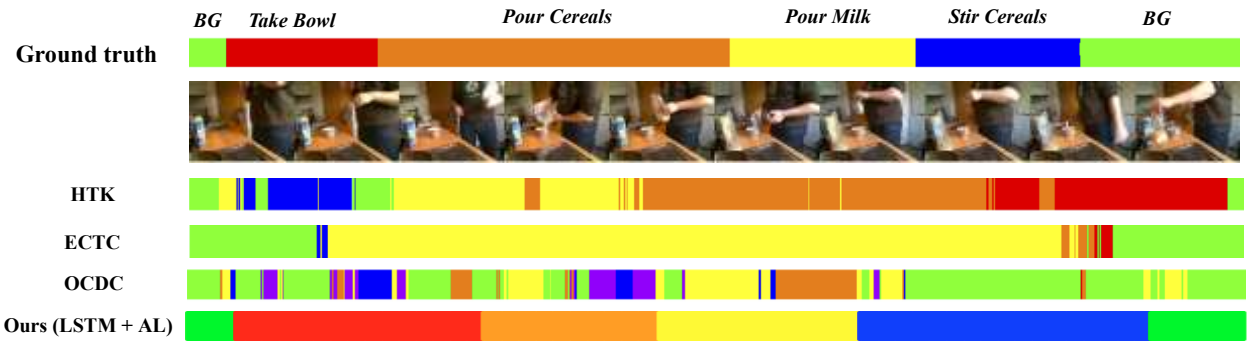

Figure 3: Illustration of the segmentation performance of the proposed approach on the Breakfast Actions Dataset on a video with ground truth *Make Cereals*. The proposed approach does not show the tendency to over-segment and provides coherent segmentation. The approach, however, shows a tendency to take longer to detect boundaries for visually similar activities.

Each activity consists of multiple sub-activities that possess visual and temporal variations according to the subject's preferences and style. Varying qualities of visual data as well as complexities such as occlusions and viewpoints increase the complexity of the temporal segmentation task.

INRIA Instructional Videos Dataset contains 150 videos of 5 different activities collected from YouTube. Each of the videos are, on average, 2 minutes long and have around 47 sub-activities. There also exists a "background activities" which consists of sequence where there does not exist a clear sub-activity that is visually discriminable. This offers a considerable challenge for approaches that are not explicitly trained for such visual features.

50 Salads Dataset is a multimodal data collected in the cooking domain. The datasets contains over four (4) hours of annotated data of 25 people preparing 2 mixed salads each and provides data in different modalities such as RGB frames, depth maps and accelerometer data for devices attached to different items such as knives, spoons and bottles to name a few. The annotations of activities are provided at different levels of granularities - high, low and eval. We use the "eval" granularity following evaluation protocols in prior works [21, 27].

4.2. Evaluation Metrics

We use two commonly used evaluation metrics for analyzing the performance of the proposed model. We used the same evaluation protocol and code as in [3, 30]. We used the Hungarian matching algorithm to obtain the one-to-one mappings between the predicted segments and the ground truth to evaluate the performance due to the unsupervised nature of the proposed approach. We use the mean over frames (MoF) to evaluate the ability of the proposed approach to temporally localize the sub-activities. We evaluate the divergence of the predicted segments from the ground truth segmentation using the Jaccard index (Intersection over Union or IoU). We also use the F1 score to evaluate the quality of the temporal segmentation. The evaluation protocol for the recognition task in Section 4.4.1 is the unit level accuracy for the 48 classes as seen in Table 3 from [19] and compared in [19, 1, 9, 16].

4.3. Ablative Studies

We evaluate different variations of our proposed approach to compare the effectiveness of each proposed component. We varied the prediction history n and the prediction error threshold Ψ. Increasing frame window tends to merge frames and smaller clusters near the event boundaries to the prior activity class due to transient increase in error. This results in higher IoU and lower MoF. Low error threshold results in over segmentation as boundary detection becomes sensitive to small changes. The number of predicted clusters decreases as the window size and threshold increases. We also trained four (4) models, with different predictor units. We trained two recurrent neural networks (RNN) as the predictor units with and without adaptive learning described in Section 3.4 indicated as *RNN + No AL* and *RNN + AL*, respectively. We also trained LSTM without adaptive learning (*LSTM + No AL*) to compare against our main model (*LSTM + AL*). We use RNNs as a possible alternative due to the short-term future predictions (1 frame ahead) required. We discuss these results next.

4.4. Quantitative Evaluation

Breakfast Actions Dataset We evaluate the performance of our full model *LSTM + AL* on the breakfast actions dataset and compare against fully supervised, weakly supervised and unsupervised approaches. We show the performance of the SVM[19] approach to highlight the importance of temporal modeling. As can be seen from Table 1, the proposed approach outperformed all unsupervised and weakly supervised approaches, and some fully supervised approaches.

It should be noted that the other unsupervised ap-