A Perceptual Prediction Framework for Self Supervised Event Segmentation

Sathyanarayanan N. Aakur
University of South Florida
Tampa, FL, USA
saakur@mail.usf.edu

Sudeep Sarkar
University of South Florida
Tampa, FL, USA
sarkar@usf.edu

Abstract

Temporal segmentation of long videos is an important problem, that has largely been tackled through supervised learning, often requiring large amounts of annotated training data. In this paper, we tackle the problem of self-supervised temporal segmentation that alleviates the need for any supervision in the form of labels (full supervision) or temporal ordering (weak supervision). We introduce a self-supervised, predictive learning framework that draws inspiration from cognitive psychology to segment long, visually complex videos into constituent events. Learning involves only a single pass through the training data. We also introduce a new adaptive learning paradigm that helps reduce the effect of catastrophic forgetting in recurrent neural networks. Extensive experiments on three publicly available datasets - Breakfast Actions, 50 Salads, and INRIA Instructional Videos datasets show the efficacy of the proposed approach. We show that the proposed approach outperforms weakly-supervised and unsupervised baselines by up to 24% and achieves competitive segmentation results compared to fully supervised baselines with only a single pass through the training data. Finally, we show that the proposed self-supervised learning paradigm learns highly discriminating features to improve action recognition.

1. Introduction

Video data can be seen as a continuous, dynamic stream of visual cues encoded in terms of coherent, stable structures called "*events*". Computer vision research has largely focused on the problem of recognizing and describing these events in terms of either labeled actions [19, 18, 19, 2, 1] or sentences (captioning) [1, 36, 35, 13, 5, 39]. Such approaches assume that the video is already segmented into atomic, stable units sharing a semantic structure such as "*throw ball*" or "*pour water*". However, the problem of temporally localizing events in untrimmed video has not been explored to the same extent as activity recognition or captioning. In this work, we aim to tackle the problem of

temporal segmentation of untrimmed videos into its constituent events in a self-supervised manner, without the need for training data annotations.
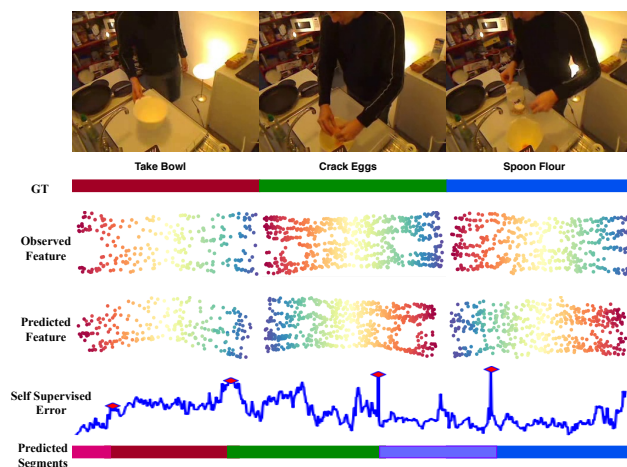
Figure 1: **Proposed Approach**: Given an unsegmented input video, we encode it into a higher level feature. We predict the feature for next time instant. A self-supervised signal based on the difference between the predicted and the observed feature gives rise to a possible event boundary.

To segment a video into its constituent *events*, we must first define the term *event*. Drawing from cognitive psychology [42], we define an event to be a "*segment of time at a given location that is perceived by an observer to have a beginning and an end*". Event segmentation is the process of identifying these beginnings and endings and their relations. Based on the level of distinction, the granularity of these events can be variable. For example, *throw ball* and *hit ball* can be events that constitute a larger, overarching event *play baseball*. Hence, each event can be characterized by a stable, internal representation that can be used to anticipate future visual features within the same event with high correlation, with increasing levels of error as the current transitions into the next. Self-supervised learning paradigms of "*predict, observe and learn*" can then be used to provide

supervision for training a computational model, typically a neural network with recurrence for temporal coherence.

We propose a novel computational model [1] based on the concept of perceptual prediction. Defined in cognitive psychology, it refers to the hierarchical process that transforms the current sensory inputs into state representations of the near future that allow for actions. Such representation of the near future enables us to anticipate sensory information based on the current event. This is illustrated in Figure 1. The features were visualized using T-SNE [23] for presentation. The proposed approach has three characteristics. It is hierarchical, recurrent and cyclical. The hierarchical nature of the proposed approach lies in the abstraction of the incoming video frames into features of lower variability that is conducive to prediction. The proposed model is also recurrent. The predicted features are highly dependent on the current and previous states of the network. Finally, the model is highly cyclical. Predictions are compared continuously to observed features and are used to guide future predictions. These characteristics are common working assumptions in many different theories of perception [26], neuro-physiology [11, 7], language processing [34] and event perception[14].

Contributions: The contributions of our proposed approach are three-fold. (1) We are, to the best of our knowledge, the first to tackle the problem of self-supervised, temporal segmentation of videos. (2) We introduce the notion of self-supervised predictive learning for active event segmentation. (3) We show that understanding the spatial-temporal dynamics of events enable the model to learn the visual structure of events for better activity recognition.

2. Related Work

Fully supervised approaches treat event segmentation as a *supervised* learning problem and assign the semantics to the video in terms of labels and try to segment the video into its semantically coherent "*chunks*", with contiguous frames sharing the same label. There have been different approaches to supervised action segmentation such as frame-based labeling using handcrafted features and a support vector machine [19], modeling temporal dynamics using Hidden Markov Models [19], temporal convolutional neural networks (TCN) [27], spatiotemporal convolutional neural networks (CNN) [21] and recurrent networks [29] to name a few. Such approaches often rely on the quantity and quality of the training annotations and constrained by the semantics captured in the training annotations, i.e., a closed world assumption.

Weakly supervised approaches have also been explored to an extent to alleviate the need for large amounts of

labeled data. The underlying concept behind weak supervision is to alleviate the need for direct labeling by leveraging accompanying text scripts or instructions as indirect supervision for learning highly discriminant features. There have been two common approaches to weakly supervised learning for temporal segmentation of videos - (1) using script or instructions for weak annotation[6, 10, 3, 24], and (2) following an incomplete temporal localization of actions for learning and inference[16, 29]. While such approaches model the temporal transitions using RNNs, they still rely on enforcing semantics for segmenting actions and hence require some supervision for learning and inference.

Unsupervised learning has not been explored to the same extent as supervised approaches, primarily because label semantics, if available, aid in segmentation. The primary approach is to use clustering as the unsupervised approach using discriminant features[4, 30]. The models incorporate a temporal consistency into the segmentation approach by using either LSTMs [4] or generalized mallows model [30]. Garcia *et al.* [12] explore the use of a generative LSTM network to segment sequences like we do, however, they handle only coarse temporal resolution in life-log images sampled as far apart as 30 seconds. Consecutive images when events change have more variability making for easier discrimination. Besides, they require an iterative training process, which we do not.

3. Perceptual Prediction Framework

In this section, we introduce the proposed framework. We begin with a discussion on the perceptual processing unit, including encoding, prediction and feature reconstruction. We continue with an explanation of the self-supervised approach for training the model, followed by a discussion on boundary detection and adaptive learning. We conclude with implementation details of the proposed approach. It is to be noted that [25] also propose a similar approach based on the Event Segmentation Theory. However, the event boundary detection is achieved using a reinforcement learning paradigm that requires significant amounts of training data and iterations and the approach has only been demonstrated on motion capture data.

3.1. Perceptual Processing

We follow the general principles outlined in the Event Segmentation Theory proposed by Zacks *et al.* [41, 42, 40]. At the core of the approach, illustrated in Figure 2 is a predictive processing platform that encodes a visual input $I(t)$ into a higher level abstraction $I'(t)$ using an encoder network. The abstracted feature is used as a prior to predict the anticipated feature $I'(t+1)$ at time $t+1$. The reconstruction or decoder network creates the anticipated feature, which is used to determine the event boundaries between successive activities in streaming, input video.
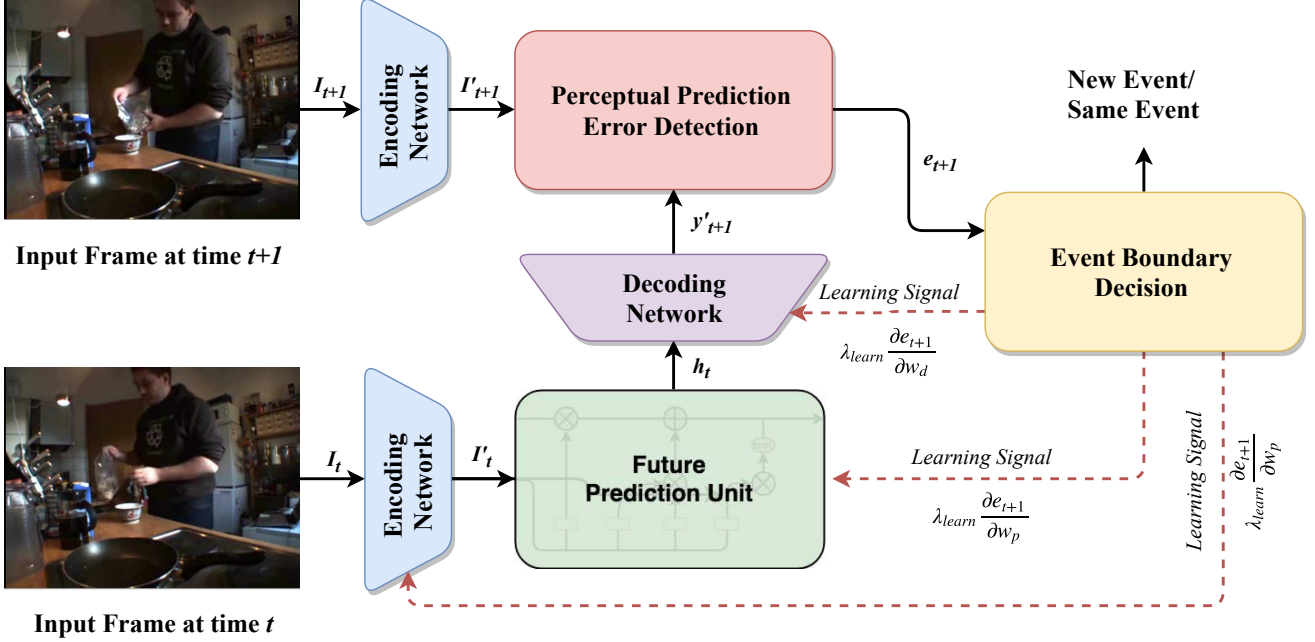
Figure 2: **Overall architecture**: The proposed approach consists of four essential components: an encoder network, a predicting unit, a decoding network, and error detection and boundary decision unit.

3.1.1 Visual Feature Encoding

We encode the input frame at each time step into an abstracted, higher level visual feature and use it as a basis for perceptual processing rather than the raw input at the pixel level (for reduced network complexity) or higher level semantics (which require training data in the form of labels). The encoding process requires learning a function $g(I(t), \omega_e)$ that transforms an input frame $I(t)$ into a higher dimensional feature space that encodes the spatial features of the input into a feature vector $I'(t)$, where ω_e is the set of learnable parameters. While the feature space can be pre-computed features such as Histogram of Optic Flow (HOF) [8], or Improved Dense Trajectories (IDT) [38], we propose the joint training of a convolutional neural network.

The prediction error and the subsequent error gradient described in Sections 3.3 and 3.4, respectively, allow for the CNN to learn highly discriminative features, resulting in higher recognition accuracy (Section 4.4.1). An added advantage is that the prediction can be made at different hierarchies of feature embeddings, including at the pixel-level, allowing for event segmentation at different granularities.

3.1.2 Recurrent Prediction for Feature Forecasting

The prediction of the visual feature at time $t+1$ is conditioned by the observation at time t, $I'(t)$, and an internal model of the current event. Formally, this can be defined by a generative model $P(I'(t+1)|\omega_p, I'(t))$, where ω_p is the

set of hidden parameters characterizing the internal state of the current observed event. To capture the temporal dependencies among *intra*-event frames and *inter*-event frames, we propose the use of a recurrent network, such as recurrent neural networks (RNN) or Long Short Term Memory Networks (LSTMs)[15]. The predictor model can be mathematically expressed as

$$
\begin{aligned}
i_t &= \sigma(W_i I'(t) + W_{hi} h_{t-1} + b_i) \quad\quad (1) \\
f_t &= \sigma(W_f I'(t) + W_{hf} h_{t-1} + b_f) \\
o_t &= \sigma(W_o I'(t) + W_{ho} h_{t-1} + b_o) \\
g_t &= \phi(W_g I'(t) + W_{hg} h_{t-1} + b_g) \\
m_t &= f_t \cdot m_{t-1} + i_t \cdot g_t \\
h_t &= o_t \cdot \phi(m_t)
\end{aligned}
$$

where σ is a non-linear activation function, the dot-operator ($\cdot$) represents element-wise multiplication, ϕ is the hyperbolic tangent function (*tanh*) and W_x and b_x represent the trained weights and biases for each of the gates. Collectively, $\{W_{hi}, W_{hf}, W_{ho}, W_{hg}\}$ and their respective biases constitute the learnable parameters ω_p.

As can be seen from Equation 1, there are four common "gates" or layers that help the prediction of the network - the input gate i_t, forget gate f_t, output gate o_t, the memory layer g_t, the memory state m_t and the event state h_t. In the proposed framework, the memory state m_t and the event state h_t are key to the predictions made by the recurrent unit. The event state h_t is a representation of the

event observed at time instant t and hence is sensitive to the observed input $I'(t)$ than the event layer, which is more persistent across events. The event layer is a gated layer, which receives input from the encoder as well as the recurrent event model. However, the inputs to the event layer are modulated by a self-supervised gating signal (Section 3.3), which is indicative of the quality of predictions made by the recurrent model. The gating allows for updating the weights quickly but also maintains a coherent state within the event.

Why recurrent networks? While convolutional decoder networks [17] and mixture-of-network models [37] are viable alternatives for future prediction, we propose the use of recurrent networks for the following reasons. Imagine a sequence of frames $I_a = (I_a^1, I_a^2, \ldots I_a^n)$ corresponding to the activity a. Given the complex nature of videos such as those in instructional or sports domains, the next set of frames can be followed by frames of activity b or c with equal probability, given by $I_b = (I_b^1, I_b^2, \ldots I_b^m)$ and $I_c = (I_c^1, I_c^2, \ldots I_c^k)$ respectively. Using a fully connected or convolutional prediction unit is likely to result in the prediction of features that tend to be the average of the two activities a and b, i.e. $I_{avg}^k = \frac{1}{2}(I_b^k + I_c^k)$ for the time k. This is not a desirable outcome because the predicted features can either be an unlikely outcome or, more probably, be outside the plausible manifold of representations. The use of recurrent networks such as RNNs and LSTMs allow for multiple futures that can be possible at time $t + 1$, conditioned upon the observation of frames until time t.

3.1.3 Feature Reconstruction

In the proposed framework, the goal of the perceptual processing unit (or rather the reconstruction network) is to reconstruct the predicted feature y'_{t+1} given a source prediction h_t, which maximizes the probability

$$p(y'_{t+1}|h_t) \propto p(h_t|y'_{t+1})\, p(y'_{t+1}) \qquad (2)$$

where the first term is the likelihood model (or translation model from NLP) and the second is the feature prior model. However, we model $\log p(y'_{t+1}|h_t)$ as a log-linear model $f(\cdot)$ conditioned upon the weights of the recurrent model ω_p and the observed feature $I'(t)$ and characterized by

$$\log p(y'_{t+1}|h_t) = \sum_{n=1}^{t} f(\omega_p, I'(t)) + log\, Z(h_t) \qquad (3)$$

where $Z(h_t)$ is a normalization constant that does not depend on the weights ω_p. The reconstruction model completes the generative process for forecasting the feature at time $t + 1$ and helps in constructing the self-supervised learning setting for identifying event boundaries.

3.2. Self-Supervised Learning

The quality of the predictions is determined by comparing the prediction from the predictor model $y'(t)$ to the observed visual feature $I'(t)$. The deviation of the predicted input from the observed features is termed as the perceptual prediction error $E_P(t)$ and is described by the equation:

$$E_P(t) = \sum_{i=1}^{n} \|I'(t) - y'(t)\|_{\ell_1}^2 \qquad (4)$$

where $E_P(t)$ is the perceptual prediction error at time t, given the predicted visual $y'(t)$ and the actual observed feature at time t, $I'(t)$. The predicted input is obtained through the inference function defined in Equation 2. The perceptual prediction error is indicative of the prediction quality and is directly correlated with the quality of the recurrent model's internal state $h(t)$. Increasingly large deviations indicate that the current state is not a reliable representation of the observed event. Hence, the gating signal serves as an indicator of event boundaries. The minimization of the perceptual prediction error serves as the objective function for the network during training.

3.3. Error Gating for Event Segmentation

The gating signal (Section 3) is an integral component in the proposed framework. We hypothesize that the visual features of successive events differ significantly at event boundaries. The difference in visual features can be minor among sub-activities and can be large across radically different events. For example, in Figure 1, we can see that the visual representation of the features learned by the encoder network for the activities *take bowl* and *crack eggs* are closer together than the features between the activities *take bowl* and *spoon flour*. This diverging feature space causes a transient increase in the perceptual prediction error, especially at event boundaries. The prediction error decreases as the predictor model adapts to the new event. This is illustrated in Figure 1. We show the perceptual prediction error (second from the bottom) and the ground truth segmentation (second from the top) for the video Make Pancake. As illustrated, the error rates are higher at the event boundaries and lower among "in-event" frames.

The unsupervised gating signal is achieved using an anomaly detection module. In our implementation, we use a low pass filter. The low pass filter maintains a relative measure of the perceptual prediction error made by the predictor module. It is a relative measure because the low pass filter only maintains a running average of the prediction errors made over the last n time steps. The perceptual quality metric, P_q, is given by:

$$P_q(t) = P_q(t-1) + \frac{1}{n}(E_P(t) - P_q(t-1)) \qquad (5)$$

where n is the prediction error history that influences the anomaly detection module's internal model for detecting event boundaries. In our experiments, we maintain n at 5. This is chosen based on the average response time of human perception, which is around 200 ms [33].

The gating signal, $G(t)$, is triggered when the current prediction error exceeds the average quality metric by at least 50%.

$$G(t) = \begin{cases} 1, & \frac{E_P(t)}{P_q(t-1)} > \psi_e \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

where $P_E(t)$ is the perceptual prediction error at time t, $G(t)$ is the value of the gating signal at time t, $P_q(t-1)$ is the prediction quality metric at time t and Ψ_e is the prediction error threshold for boundary detection. For optimal prediction, the perceptual prediction error would be very high at the event boundary frames and very low at all within-event frames. In our experiments, Ψ_e is set to be 1.5.

In actual, real-world video frames, however, there exist additional noise in the form of occlusions and background motion which can cause some event boundaries to have a low perceptual prediction error. In that case, however, the gating signal would continue to be low and become high when there is a transient increase in error. This is visualized in Figure 1. It can be seen that the perceptual errors were lower at event boundaries between activities *take bowl* and *crack eggs* in a video of ground truth *make pancakes*. However, the prediction error increases radically soon after the boundary frames, indicating a new event. Such cases could, arguably, be attributed to conditions when there are lesser variations in the visual features at an event boundary.

3.4. Adaptive Learning for Plasticity

The proposed training of the prediction module is particularly conducive towards overfitting since we propagate the perceptual prediction error at every time step. This introduces severe overfitting, especially in the prediction model. To allow for some plasticity and avoid catastrophic forgetting in the network, we introduce the concept of adaptive learning. This is similar to the learning rate schedule, a commonly used technique for training deep neural networks. However, instead of using predetermined intervals for changing the learning rates, we propose the use of the gating signal to modulate the learning rate. For example, when the perceptual prediction rate is lower than the average prediction rate, the predictor model is considered to have a good, stable representation of the current event. Propagating the prediction error, when there is a good representation of the event can lead to overfitting of the predictor model to that particular event and does not help generalize. Hence, we propose lower learning rates for time steps when there are negligible prediction error and a relatively higher (by a magnitude of 100) for when there is

higher prediction error. Intuitively, this adaptive learning rate allows the model to adapt much quicker to new events (at event boundaries where there are likely to be higher errors) and learn to maintain the internal representation for within-event frames.

Formally, the learning rate is defined as the result of the adaptive learning rule defined as a function of the perceptual prediction error defined in Section 3.2 and is defined as

$$\lambda_{learn} = \begin{cases} \Delta_t^- \lambda_{init}, & E_P(t) > \mu_e \\ \Delta_t^+ \lambda_{init}, & E_P(t) < \mu_e \\ \lambda_{init}, & otherwise \end{cases} \quad (7)$$

where Δ_t^-, Δ_t^+ and λ_{init} refer to the scaling of the learning rate in the negative direction, positive direction and the initial learning rate respectively and $\mu_e = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} E_P \, dE_P$. The learning rate is adjusted based on the quality of the predictions characterized by the perceptual prediction error between a temporal sequence between times t_1 and t_2, typically defined by the gating signal.. The impact of the adaptive changes to the learning rate is shown in the quantitative evaluation Section 4.4, where the adaptive learning scheme shows improvement of up to 20% compared to training without the learning scheme.

3.5. Implementation Details

In our experiments, we use a VGG-16 [31] network pretrained on ImageNet as our hierarchical, feature encoder module. We discard the final layer and use the second fully connected layer with 4096 units as our encoded feature vector for a given frame. The feature vector is then consumed by a predictor model. We trained two versions, one with an RNN and the other with an LSTM as our predictor models. The LSTM model used is the original version proposed by [15]. Finally, the anomaly detection module runs an average low pass filter described in Section 3.3. The initial learning rate described in Section 3.4 is set to be 1×10^{-6}. The scaling factors Δ_t^- and Δ_t^+ are set to be 1×10^{-2} and 1×10^{-3}, respectively. The training was done on a computer with one Titan X Pascal.

4. Experimental Evaluation

4.1. Datasets

We evaluate and analyze the performance of the proposed approach on three large, publicly available datasets - Breakfast Actions [19], INRIA Instructional Videos dataset[3] and the 50 Salads dataset [32]. Each dataset offers a different challenge to the approach allow us to evaluate its performance on a variety of challenging conditions.

Breakfast Actions Dataset is a large collection of 1,712 videos of 10 breakfast activities performed by 52 actors.
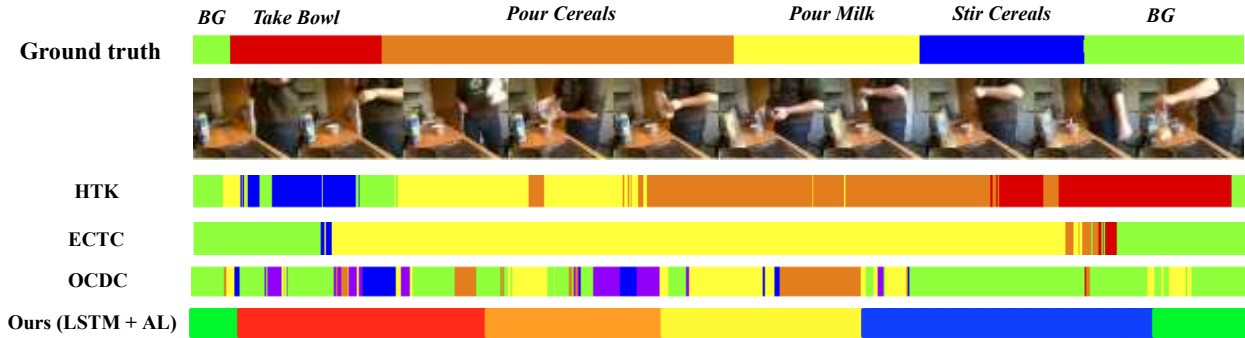
Figure 3: Illustration of the segmentation performance of the proposed approach on the Breakfast Actions Dataset on a video with ground truth *Make Cereals*. The proposed approach does not show the tendency to over-segment and provides coherent segmentation. The approach, however, shows a tendency to 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.

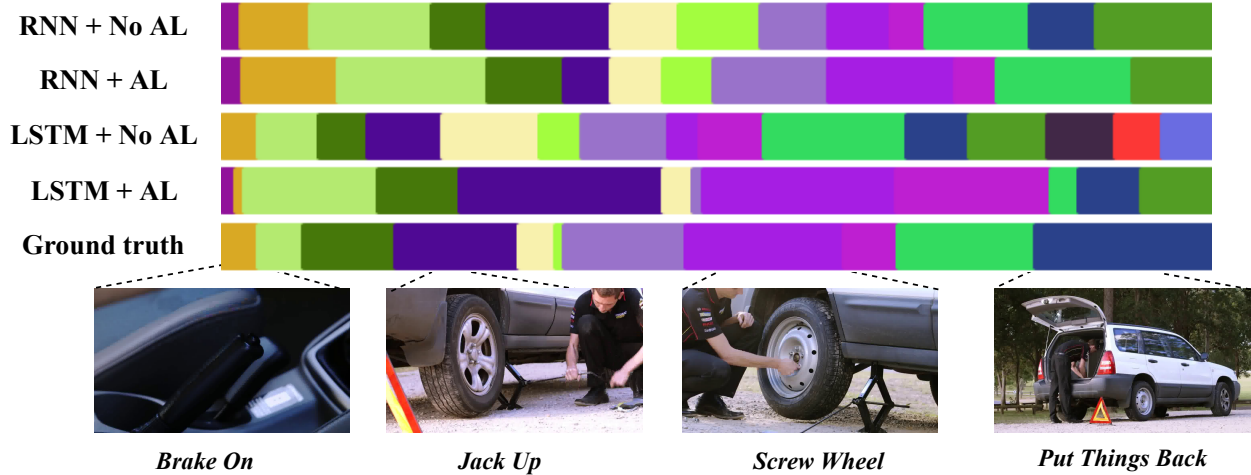It should be noted that the other unsupervised ap-

Figure 4: **Ablative Studies**: Illustrative comparison of variations of our approach, using RNNs and LSTMs with and without adaptive learning on the INRIA Instructional Videos Dataset on a video with ground truth *Change Tire*. It can be seen that complex visual scenes with activities of shorter duration pose a significant challenge to the proposed framework and cause fragmentation and over segmentation. However, the use of adaptive learning helps alleviate this to some extent. Note: Temporal segmentation time lines are shown without the background class for better visualization.

Supervision	Approach	MoF	IoU
Full	SVM [19]	15.8	-
	HTK(64)[20]	56.3	-
	ED-TCN[27]	43.3	42.0
	TCFPN[10]	52.0	**54.9**
	GRU[29]	**60.6**	-
Weak	OCDC[6]	8.9	23.4
	ECTC[16]	27.7	-
	Fine2Coarse[28]	33.3	**47.3**
	TCFPN + ISBA[10]	**38.4**	40.6
None	KNN+GMM[30]	34.6	**47.1**
	Ours (LSTM + AL)	**42.9**	46.9

Table 1: Segmentation Results on the Breakfast Action dataset. MoF refers to the Mean over Frames metric and IoU is the Intersection over Union metric.

proach [30], requires the number of clusters (from ground truth) to achieve the performance whereas our approach does not require such knowledge and is done in a streaming fashion. Additionally, the weakly supervised methods [16, 28, 10] require both the number of actions as well as an ordered list of sub-activities as input. ECTC [16] is based on discriminative clustering, while OCDC [6] and Fine2Coarse [28] are also RNN-based methods.

50 Salads Dataset We also evaluate our approach on the 50 Salads dataset, using only the visual features as input. We report the Mean of Frames (MoF) metric for fair comparison. As can be seen from Table 2, the proposed approach significantly outperforms the other unsupervised approach, improving by over 11%. We also show the perfor-

mance of the frame-based classification approaches VGG and IDT [21] to show the impact of temporal modeling. It should be noted that the fully supervised approaches re-

Supervision	Approach	MoF
Full	VGG**[21]	7.6%
	IDT**[21]	54.3%
	S-CNN + LSTM[21]	66.6%
	TDRN[22]	68.1%
	ST-CNN + Seg[21]	72.0%
	TCN[27]	**73.4%**
None	LSTM + KNN[4]	54.0%
	Ours (LSTM + AL)	**60.6%**

Table 2: Segmentation Results on the 50 Salads dataset, at granularity '*Eval*'. **Models were intentionally reported without temporal constraints for ablative studies.

quired significantly more training data - both in the form of labels as well as training epochs. Additionally, the TCN approach [27] uses the accelerometer data as well to achieve the state-of-the-art performance of 74.4%

INRIA Instructional Videos Dataset: Finally, we evaluate our approach on the INRIA Instructional Videos dataset, which posed a significant challenge in the form of high amounts of background (noise) data. We report the F1 score for fair comparison to the other state-of-the-art approaches. As can be seen from Table 3, the proposed model outperforms the other unsupervised approach [30] by 23.3%, the weakly supervised approach [6] by 24.8% and has competitive performance to the fully supervised approaches[24, 3, 30].

Supervision	Approach	F1
Full	HMM + Text [24]	22.9%
	Discriminative Clustering[3]	41.4%
	KNN+GMM[30] + GT	**69.2%**
Weak	OCDC + Text Features [6]	28.9%
	OCDC [6]	**31.8%**
None	KNN+GMM[30]	32.2%
	Ours (RNN + No AL)	25.9%
	Ours (RNN + AL)	29.4%
	Ours (LSTM + No AL)	36.4%
	Ours (LSTM + AL)	**39.7%**

Table 3: Segmentation Results on the INRIA Instructional Videos dataset. We report F1 score for fair comparison.

We also evaluate the performance of the models with and without adaptive learning. It can be seen that long term temporal dependence captured by LSTMs is significant, especially due to the long durations of activities in the dataset. Additionally, the use of adaptive learning has a significant improvement in the segmentation framework, improving the performance by 9% and 11% for the RNN-based model and the LSTM-based model respectively, indicating a reduced overfitting of the model to the visual data.

4.4.1 Improved Features for Action Recognition

To evaluate the ability of the network to learn highly discriminative features for recognition, we evaluated the performance of the proposed approach in a recognition task. We use the model pretrained on the segmentation task on the Breakfast Actions dataset and use the hidden layer of the LSTM as input to a fully connected layer and use cross entropy to train the model. We also trained another network with the same structure - VGG16 + LSTM without the pretraining on the segmentation task to compare the effect of the features learned using self-supervision. As can

Approach	Precision
HCF + HMM [19]	14.90%
HCF + CFG + HMM [19]	31.8%
RNN + ECTC [16]	35.6%
RNN + ECTC (Cosine) [16]	36.7%
HCF + Pattern Theory [9]	38.6%
HCF + Pattern Theory + ConceptNet[1]	42.9%
VGG16 + LSTM	33.54%
VGG16 + LSTM + Predictive Features(AL)	37.87%

Table 4: Activity recognition results on Breakfast Actions dataset. HCF and AL refer to handcrafted features and Adaptive Learning, respectively.

be seen from Table 4, the use of self-supervision to pretrain the network prior to the recognition task improves the recognition performance of the network and has comparable performance to the other state-of-the-art approaches. It improves the recognition accuracy by 13.12% over the network without predictive pretraining.

4.5. Qualitative Evaluation

Through the predictive, self supervised framework, we are able to learn the sequence of visual features in streaming video. We visualize the segmentation performance of the proposed framework on the Breakfast Actions Dataset in Figure 3. It can be seen that the proposed approach has high temporal coherence and does not suffer from over segmentation, especially when the segments are long. Long activity sequences allow the model to learn from observation by providing more samples of "*intra*-event" samples. Additionally, it can be seen that weakly supervised approaches like OCDC[6] and ECTC[16] suffer from over segmentation and intra-class fragmentation. This could arguably be attributed to the fact that they tend to enforce semantics, in the form of weak ordering of activities in the video regardless of the changes in visual features. Fully supervised approaches, such as HTK[20] perform better, especially due to the ability to assign semantics to visual features. However, they are also affected by unbalanced data and dataset shift, as can be seen in Figure 3 where the background class was segmented into other classes.

We also qualitatively evaluated the impact of adaptive learning and long term temporal memory in Figure 4, where the performance of the alternative methods described in Section 4.4. It can be seen that the use of adaptive learning during training allows the model to not overfit to any single class' intra-event frames and help generalize to other classes regardless of amount of training data. It is not to say that the problem of unbalanced data is alleviated, but the adaptive learning *does* help to some extent. It is interesting to note that the LSTM model tends to over-segment when not trained with adaptive learning, while the RNN-based model does not suffer from the same fate.

5. Conclusion

We demonstrate how a self-supervised learning paradigm can be used to segment long, highly complex visual sequences. There are key differences between our approach and fully supervised or weakly supervised approaches, including classical ones such as DBMs and HMMs. At a high level, our approach is unsupervised and does not require labeled data for training. The predictive error serves as supervision for training the framework. The other major aspect is that our approach requires only a single pass through the training data. Hence, the training time is very low. The experimental results demonstrate the robustness, high performance, and the generality of the approach on multiple real world datasets.

References

[1] Sathyanarayanan Aakur, Fillipe DM de Souza, and Sudeep Sarkar. Going deeper with semantics: Exploiting semantic contextualization for interpretation of human activity in videos. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019. 1, 6, 8

[2] Sathyanarayanan N. Aakur, Fillipe DM de Souza, and Sudeep Sarkar. Towards a knowledge-based approach for generating video descriptions. In *Conference on Computer and Robot Vision (CRV)*. Springer, 2017. 1

[3] Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4575–4583, 2016. 2, 5, 6, 7, 8

[4] Bharat Lal Bhatnagar, Suriya Singh, Chetan Arora, CV Jawahar, and KCIS CVIT. Unsupervised learning of deep feature representation for clustering egocentric actions. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1447–1453. AAAI Press, 2017. 2, 7

[5] Yi Bin, Yang Yang, Fumin Shen, Xing Xu, and Heng Tao Shen. Bidirectional long-short term memory for video description. In *ACM Conference on Multimedia (ACM MM)*, pages 436–440. ACM, 2016. 1

[6] Piotr Bojanowski, Rémi Lajugie, Francis Bach, Ivan Laptev, Jean Ponce, Cordelia Schmid, and Josef Sivic. Weakly supervised action labeling in videos under ordering constraints. In *European Conference on Computer Vision (ECCV)*, pages 628–643. Springer, 2014. 2, 7, 8

[7] Gail A Carpenter and Stephen Grossberg. *Adaptive resonance theory*. Springer, 2016. 2

[8] Rizwan Chaudhry, Avinash Ravichandran, Gregory Hager, and René Vidal. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1932–1939. IEEE, 2009. 3

[9] Fillipe DM de Souza, Sudeep Sarkar, Anuj Srivastava, and Jingyong Su. Spatially coherent interpretations of videos using pattern theory. *International Journal on Computer Vision (IJCV)*, pages 1–21, 2016. 6, 8

[10] Li Ding and Chenliang Xu. Weakly-supervised action segmentation with iterative soft boundary assignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 7

[11] Joaquin M Fuster. The prefrontal cortex and its relation to behavior. In *Progress in brain research*, volume 87, pages 201–211. Elsevier, 1991. 2

[12] Ana Garcia del Molino, Joo-Hwee Lim, and Ah-Hwee Tan. Predicting visual context for unsupervised event segmentation in continuous photo-streams. In *ACM Conference on Multimedia (ACM MM)*, pages 10–17. ACM, 2018. 2

[13] Zhao Guo, Lianli Gao, Jingkuan Song, Xing Xu, Jie Shao, and Heng Tao Shen. Attention-based lstm with semantic consistency for videos captioning. In *ACM Conference on Multimedia (ACM MM)*, pages 357–361. ACM, 2016. 1

[14] Catherine Hanson and Stephen José Hanson. Development of schemata during event parsing: Neisser's perceptual cycle as a recurrent connectionist network. *Journal of Cognitive Neuroscience*, 8(2):119–134, 1996. 2

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 3, 5

[16] De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. Connectionist temporal modeling for weakly supervised action labeling. In *European Conference on Computer Vision (ECCV)*, pages 137–153. Springer, 2016. 2, 6, 7, 8

[17] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. In *Neural Information Processing Systems*, pages 667–675, 2016. 4

[18] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1725–1732, 2014. 1

[19] Hilde Kuehne, Ali Arslan, and Thomas 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. 1, 2, 5, 6, 7, 8

[20] Hilde Kuehne, Juergen Gall, and Thomas Serre. An end-to-end generative framework for video segmentation and recognition. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–8. IEEE, 2016. 7, 8

[21] Colin Lea, Austin Reiter, René Vidal, and Gregory D Hager. Segmental spatiotemporal cnns for fine-grained action segmentation. In *European Conference on Computer Vision (ECCV)*, pages 36–52. Springer, 2016. 2, 6, 7

[22] Peng Lei and Sinisa Todorovic. Temporal deformable residual networks for action segmentation in videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6742–6751, 2018. 7

[23] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008. 2

[24] Jonathan Malmaud, Jonathan Huang, Vivek Rathod, Nick Johnston, Andrew Rabinovich, and Kevin Murphy. What's cookin'? interpreting cooking videos using text, speech and vision. *arXiv preprint arXiv:1503.01558*, 2015. 2, 7, 8

[25] Katherine Metcalf and David Leake. Modelling unsupervised event segmentation: Learning event boundaries from prediction errors. In *CogSci*, 2017. 2

[26] Ulric Neisser. Cognitive psychology new york: Appleton-century-crofts. *Google Scholar*, 1967. 2

[27] Colin Lea Michael D Flynn René and Vidal Austin Reiter Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 2, 6, 7

[28] Alexander Richard and Juergen Gall. Temporal action detection using a statistical language model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3131–3140, 2016. 7

[29] Alexander Richard, Hilde Kuehne, and Juergen Gall. Weakly supervised action learning with rnn based fine-to-coarse

modeling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 3, 2017. 2, 7

[30] Fadime Sener and Angela Yao. Unsupervised learning and segmentation of complex activities from video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 6, 7, 8

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[32] Sebastian Stein and Stephen J McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 729–738. ACM, 2013. 5

[33] Simon Thorpe, Denis Fize, and Catherine Marlot. Speed of processing in the human visual system. *Nature*, 381(6582):520, 1996. 5

[34] Teun Adrianus Van Dijk, Walter Kintsch, and Teun Adrianus Van Dijk. *Strategies of discourse comprehension*. Academic Press New York, 1933. 2

[35] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4534–4542, 2015. 1

[36] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*, 2014. 1

[37] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating visual representations from unlabeled video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 98–106, 2016. 4

[38] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3551–3558, 2013. 3

[39] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. Describing videos by exploiting temporal structure. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4507–4515, 2015. 1

[40] Jeffrey M Zacks and Khena M Swallow. Event segmentation. *Current Directions in Psychological Science*, 16(2):80–84, 2007. 2

[41] Jeffrey M Zacks and Barbara Tversky. Event structure in perception and conception. *Psychological bulletin*, 127(1):3, 2001. 2

[42] Jeffrey M Zacks, Barbara Tversky, and Gowri Iyer. Perceiving, remembering, and communicating structure in events. *Journal of Experimental Psychology: General*, 130(1):29, 2001. 1, 2