NJICE –National Journal on Information and Communication Engineering
ISSN: 2231-2099 - Volume 6 Issue 1
Apr-Jun 2016 Pages 20-28

Forecast of Long span Complex Human Movement in Videos

Prof.Shweta M Nirmanik
Assistant Professor
Deparment of Computer Science & Engineering
R,T,E Society's Rural Engineering College Hulkoti-582201

Abstract

Human activities are naturally structured as hierarchies unrolled overtime. For action prediction, temporal relations in event sequences are widely exploited by current methods while their semantic coherence across different levels of abstraction has not been well explored. In this work we model the hierarchical structure of human activities in videos and demonstrate the power of such structure in action prediction. We propose Hierarchical Encoder Refresher-Anticipator, a multi-level neural machine that can learn the structure of human activities by observing a partial hierarchy of events and roll-out such structure into a future prediction in multiple levels of abstraction. We also introduce a new coarse-to-fine action annotation on the Breakfast Actions videos to create a comprehensive, consistent, and cleanly structured video hierarchical activity dataset. Through our experiments, we examine and rethink the settings and metrics of activity prediction tasks toward unbiased evaluation of prediction systems, and demonstrate the role of hierarchical modeling toward reliable and detailed long-term action forecasting.

1 Introduction

An AI agent that shares the world with us needs to efficiently anticipate human activities to be able to react to them. Moreover, the ability to anticipate human activities is a strong indicator of the competency in human behavior understanding by artificial intelligence systems. While video action recognition and short-term prediction have made much progress, reliable long-term anticipation of activities remains challenging as it requires deeper understanding of the action patterns.

The most successful methods for activity prediction rely on modeling the continuity of action sequences to estimate future occurrence by neural networks. However, these networks only consider the sequential properties of the action sequence which tends to fade and entice error accumulation in far-term. This issue suggests exploring the abstract structure of actions that spans over the whole undertaking of the task. One intuitive way to approach this path is to follow the natural human planning process that starts with high level tasks then proceeds to more refined sub-tasks and detailed actions. An example of such structure in an activity is shown in Fig. 1. Our quest is to build a neural

machine that can learn to explore such structures by observing a limited section of the video and extrapolate the activity structure into the future for action prediction.

Figure 1: Illustration of a two-level structure of activity "have dinners" and a prediction task.

NJICE –National Journal on Information and Communication Engineering
ISSN: 2231-2099 - Volume 6 Issue 1
Apr-Jun 2016 Pages 20-28

We realize this vision by designing a neural architecture called Hierarchical Encoder-Refresher- Anticipator (HERA) for activity prediction. HERA consists of three sub-networks that consecutively encode the past, refresh the transitional states, and decode the future until the end of the overall task. The specialty of these networks is that their layers represent semantic levels of the activity hierarchy, from abstract to detail. Each of them operates on its

own clock while sending its state to parent layer and laying out plans for its children.

This model can be trained end-to-end and learn to explore and predict the hierarchical structure of new video sequences. We demonstrate the effectiveness of HERA in improved long-term predictions, increased reliability in predicting unfinished activities, and effective predictions of activities at different levels of granularity.

To promote further research in hierarchical activity structures, we also introduce a new hierarchical action annotation to the popular Breakfast Actions dataset. These annotations contain two-level action labels that are carefully designed to reflect the clean hierarchy of actions following natural human planning. In numbers, it includes 25,537 annotations in two levels on 1,717 videos spanning 77 hours. Once publicly released, this dataset will provide a key data source to support advancing deep understanding into human behaviors with potential applications in detection, segmentation and prediction.

2 Related works

For prediction of actions in videos, the most popular approach is to predict the temporal action segments, by jointly predicting the action labels and their lengths. Recent advances in this front include Farha et al. where random prediction points are used with the RNN/CNN-like model. Moving away from recurrent networks which tend to accumulate errors, Ke et al. used time point as the conditioning factor in one-shot prediction approach with the trade-off in high prediction cost and sparse predictions. While these methods work relatively well in near-term, when the actions are predicted farther into the future, uncertainty prevents them from having reliable results. Variational methods manage uncertainty by using probabilistic modeling to achieve more robust estimation of inter-arrival time and action length.

As an action is highly indicative of the next action, Miech et al. proposed a model that is a convex combination of a “predictive” model and a “transitional” model. A memory-based approach network was proposed by Gammulle et al. in which two streams with independent memories analyze visual and label features to predict the next action.

The hierarchy of activities can be considered in atomic scales where small movements constitute an action. Early works investigated the hierarchy of activity through layered HMM, layered CRF, and linguistic-like grammar. More recent works favor neural networks due to their strong inductive properties. For hierarchy, Recurrent Neural Networks (RNN) can be stacked up, but stacking ignores the multi-clock nature of a hierarchy unrolled over time. In a hierarchical RNN with asynchronous clocks was used to model the temporal point processes of activity but the information only passes upward and multi-level semantics of events are not explored. The idea of multi clocks was also explored by Hihi and Bengio and Koutnik et al. The drawback of these methods is that the periods of the clock must be manually defined, which is not adaptive to data structure at hand. Chung et al. addressed this problem with a hierarchical multi-scale RNN (HM-RNN), which automatically learns the latent hierarchical structure. This idea has been extended with attention mechanism for action recognition. Our hierarchical modeling shares the structure exploration functionality with these works but is significantly

NJICE –National Journal on Information and Communication Engineering
ISSN: 2231-2099 - Volume 6 Issue 1
Apr-Jun 2016 Pages 20-28

different in the ability to learn the semantic-rich structures where layers of hierarchy are associated with levels of activity abstraction. In particular, in comparison with Clock-work RNN (CW-RNN), HERA shares the fact that units can update at different rates, but HERA is significantly different to CW-RNN in separating the levels of RNN with distinctive associated semantics. HERA also allows RNN units to control their own clocks and theirinteractions with other units.

3 Learning to abstract and predict humanactions

Proposed Approach

The Basic steps in out proposed approach are as follows:

Training:

Our method takes all possible small atomic action units (Actionlet) such as washing vegetable, cutting vegetable etc. for training and models the relationship between partially observed video and trained actionlet at testing phase. Main steps in model learning are as follows:

I. Feature Detection and Extraction: For each single atomic action unit we first extracted dense Histogram of Optical Flow (HOF), Histogram of Gradient (HOG), Motion Boundary Histogram (MBH) features.

II. Dimensionality reduction: These densely extracted descriptors are high dimensional. So we use Random projection for reducing the dimensionality of feature vector

III. Dictionary learning: Instead of common bag of word approach, we use sparse dictionary learning. Given a set of input vector for each training video, the over complete dictionary is leaned and corresponding sparse representation is obtained using densely extracted features Class specific dictionaries are learned by solving sparse approximation problem using K-SVD algorithm .

Testing:

I. Temporal segmentation: Given a partial video consisting of complex long duration activity, first step is to temporally segment the given video such that each segment of video consists of meaningful atomic action. This step is carried out using Super frame segmentation proposed in. The idea is to find the boundaries in video where significant changes in motion occur, and then cut the video accordingly into multiple segments.

II. Feature Detection and Extraction: For each single atomic action unit in testing video extract dense HOG, HOF and MBH features.

III. Classification: The class of each observed segment is recognized using learned dictionary and Random Sample Reconstruction (RSR).

IV. Dynamic Prediction: Using the class label for each local segment of observed video, predict the global class label for unobserved video by computing Maximum Aposteriori probability

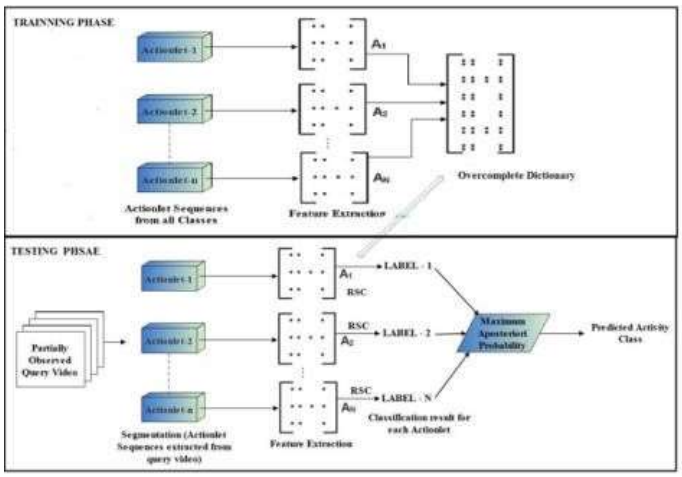

Fig. 2: Overall Pipeline of Proposed Approach

Problem formulation

NJICE –National Journal on Information and Communication Engineering
ISSN: 2231-2099 - Volume 6 Issue 1
Apr-Jun 2016 Pages 20-28

We formalize activity hierarchy H of L levels of a human performing a task observable in a video as H $= \{A^l\}_{l=1,2,\ldots,L}$ where each level A^l is a sequence of indexed actions:

$Al = \{(xl\ \ dl)\}k=1,2,\ldots..nl$ ---------- (1)

Here, x^l represents the label of k-th action at the l-th level, d^l is its relative duration calculated as its portion of the parent activity, and nl indicates the number of actions at level l. Each action $(x^l\ d^l)$ is associated with a subsequence of finer actions at level $l+1$, and the latter are called children actions of the former. Any children subsequence is constrained to exclusively belong to.

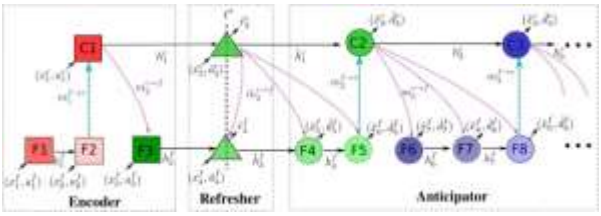

Figure 3: The Hierarchical Encoder-Refresher- Anticipator (HERA) architecture realized in a particular event sequence similar to the one in Fig. 1. Square blocks are Encoder GRU cells, while triangles and circles are those of Refresher and Anticipator, respectively. Color shades indicate cells processing different activity families, e.g., the first coarse cell (red C1) and its two children (fading red F1 and F2) process the first activity family $\{(x^c\ d^c), (x^f\ d^f), (x^f\ d^f)\}$ .The prediction point t^* Happens at the middle of $(x^c\ d^c)$ and$(x^f\ d^f)$. Black arrows indicate recurrent links while those in pink and cyan are for downward and upward messages, respectively. For visual clarity, optional prediction outputs of Encoder cell and feedback inputs of Anticipator cell are omitted.

In the special case of a hierarchy with two levels, members of the first level represent coarse activities, and those at the second level are called fine actions. In this case, we will extend the notation to use the level indices c - for coarse and f - for fine in place of numeric indices l = 1 and l = 2. An example of a two- level hierarchy is shown in Fig. 1, where for a task of have-<dinner>, the first coarse activity <prepare>-

Under this structure, the prediction problem is formed when the hierarchy of activities is interrupted at a certain time $t*$ indicating the point where observation ends. At this time, at every level we have finished events, unfinished events, and the task is to predict events yet to start. The given observation includes the labels and lengths of the finished events, and the labels and partial lengths of the unfinished ones. Thus the task boils down to estimating the remaining lengths of the unfinished events, and all details of the remaining events.

3.2 Hierarchical Encoder-Refresher Anticipator

We design HERA (Fig. 2) to natively handle the hierarchical structure of observation and extend such structure to prediction. HERA has three components: the Encoder, the Refresher, and the Anticipator. The Encoder creates a multilevel representation of the observed events which is used by the Refresher and Anticipator to roll-out in a similar manner. The Encoder and Anticipator share the same hierarchical model design for cross-level interaction which we detail next. Modeling activity hierarchy. The Encoder and Anticipator share an identical architecture of two layers of recurrent neural units (RNN) which are chosen to be based on Gated Recurrent Units (GRU). The upper layer models the dynamics of coarse activities:

$hic = GRU\ ([(xic, aic), mif\rightarrow c], hi-1c)$ ---------- (2)

The first input to the unit includes a tuple of coarse label x_{ic} and accumulated duration $a_{ic} = \Sigma d_{kcik=1}$ both x_{ic} and a_{ic} are encoded using a random embedding matrix. At the Anticipator, these inputs are feedback from the previous prediction step. The second input $m_{if\rightarrow c}$ is the upward message that will be discussed later.

The lower layer is another RNN that is triggered to start following the parent's operation:

$hjf = GRU\ ([(xjf, ajf), mic\rightarrow f], hj-1f)$ ---------- (3)

Where the proportional accumulated duration a_{jf} is calculated within the parent activity. By design, the two layers are asynchronous (i.e. the layers update their hidden state independently and whenever fit) as coarse activities happen sparser than fine actions. A key feature of HERA is the way it connects these two asynchronous concurrent processes in a consistent hierarchy by using the cross level messages. The downward message $m_{ic\rightarrow f}$ (pink arrows in Fig.2) provides instructions from the previous coarse cell

NJICE –National Journal on Information and Communication Engineering
ISSN: 2231-2099 - Volume 6 Issue 1
Apr-Jun 2016 Pages 20-28

to the current fine cells. This message contains the previous coarse hidden state $h_{i-1c}$and can optionally contain the parent's predicted label x_{ic}. The upward message $m_{ic\to f}$(cyan arrows) to a coarse node i from its children contain the information about the detail roll-out in the fine actions. It is implemented as the hidden state of the last child.

3.3 Data annotation

To support the problem structure formulated above we reannotated the Breakfast Actions videos , which is the largest multi-level video activity dataset publicly available. This dataset contains footage of 52 people preparing 10 distinct breakfast-related dishes, totaling 1,717 videos. It originally contains fine- and coarse-level annotations of the actions but the hierarchy is incoherent (inconsistent semantic abstraction), incomplete (only 804 of the videos have fine-level annotations), and statistically weak (many fine actions are less than a few frames). We employed two annotators working independently on all 1,717 videos and one verifier who checked the consistency of the annotations. Following the hierarchy definition in Sec. 3.1, we annotated a two-level hierarchy of coarse activities and fine actions. Each label of activity or action follows the format of <verb-noun> where verbs and nouns are selected from a predefined vocabulary.

The two vocabulary sets were built by a pilot round of annotation. The coarse activities can share the fine action labels. For instance, <add-salt> fine action label can be used for many coarse activities including <make-salad>, <fry-egg>, and <make-sandwich>. In actual annotation, we have 30 <verb-noun> pairs for coarse activities and 140 for fine actions that are active. The new annotation resulted in a total of 25,537 label-duration annotations with 6,549 at the coarse level and 18,988 at the fine level. We call the new annotation Hierarchical Breakfast dataset and it is available for download2, alongside the source code for HERA.

3.4 Metrics

Recent action prediction works [7, 10] widely used mean-over-class (MoC) as the key performance metric. However, MoC is susceptible to bias in class imbalance which exists in action prediction datasets. More importantly, as any framebased metrics, it merits any correctly predicted frames even when the predicted segments are mostly unaligned due to under- or over-segmentation. We verified these conceptual problems by setting up an experiment (detailed in Sec. 4) using an under-segmenting dummy predictor that takes advantage of the flaw of the metric and win over state-of-the-art methods on many settings. We call our dummy predictor "under-segmenting" because it predicts that the future consists simply of one single long action.

In the search for better metrics, we examined options including the segmental edit distance, the mean-over-frame (MoF), and the F1@k. Among them, we learned that the most suitable metric for the purpose of action predictionis the F1@k for its robustness to variation in video duration and minor shifts caused by annotation errors. Furthermore, it penalizes both over- and undersegmentations such as from our dummy predictor. This metric was previously used for temporal detection and segmentation [16]. Applied to the prediction task, we first calculate the intersection over union (IoU) of the predicted segments with the ground-truth. Any overlapping pair with IoU surpassing the chosen threshold $0 < k < 1$ is counted as correct when contributing to the final $F1 = 2\times Prec\times Recall/Prec+Recall$.

4. Experimental Results

We have tested our approach on two dataset MHOI and MPPI cooking dataset.

Results on MHOI Dataset:

MHOI is the daily activity dataset such as "answering a phone call", "drinking tea" etc. In each class of activity human interacts with some object. These dataset activities are short duration activities consists of around 2 to 4 atomic actions (actionlet) such as grabbing the object, putting it back etc.

There are total 6 different types of activities each of which is performed by 8 to 10 subject

NJICE –National Journal on Information and Communication Engineering
ISSN: 2231-2099 - Volume 6 Issue 1
Apr-Jun 2016 Pages 20-28

NJICE –National Journal on Information and Communication Engineering
ISSN: 2231-2099 - Volume 6 Issue 1
Apr-Jun 2016 Pages 20-28

Methods	Training samples used	Observation Ration				
		20%	40%	60%	80%	100%
Integral BOW	Leave-One-Out	0.32	0.41	0.42	0.50	0.52
Dynamic BOW	Leave-One-Out	0.40	0.47	0.50	0.55	0.52
BOW+SVM	Leave-One-Out	0.30	0.41	0.41	0.39	0.39
HMM	Leave-One-Out	0.23	0.38	0.56	0.47	0.43
Action Only Model	Leave-One-Out	0.37	0.37	0.63	0.65	0.65
Our Approach	70 percent	0.35	0.40	0.66	0.68	0.68

Table 2: Comparison with State-of-The-Art on MHOI Dataset

Fig. 3: Sample Actionlets From MHOI Dataset

Fig. 4: Sample Actionlets from MPPI Dataset

Results on MPPI Dataset:

NJICE –National Journal on Information and Communication Engineering
ISSN: 2231-2099 - Volume 6 Issue 1
Apr-Jun 2016 Pages 20-28

MPPI dataset is the cooking activity dataset such as "making a salad", "making a sandwich" etc.
In each class of activity human interacts with some object.These dataset activities are long duration complex activities consists of around 20 to 125 atomic actions (actionlet) such as cut slices, pour, or spice. There are total 65 different activities (actionlets) performed by various actors. There are 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

Methods	Training samples used	Observation Ration				
		20%	40%	60%	80%	100 %
HMM	Leave-One-Out	0.49	0.53	0.60	0.62	0.65
Action Only Model	Leave-One-Out	0.58	0.59	0.64	0.64	0.66
Our Approach	70 percent	0.52	0.62	0.66	0.68	0.70

Table 3: Comparison with State-of-The-Art on MPPI Dataset

5 Conclusions

We have introduced HERA (Hierarchical Encoder-Refresher-Anticipator), a new hierarchical neural network for modeling and predicting the long-term multilevel action dynamics in videos. To promote further research we re-annotated from scratch 1,717 videos in the Breakfast Actions dataset, creating a new and complete semantically coherent annotation of activity hierarchy, which we named Hierarchical Breakfast. We also reassessed the commonly used MoC metric in action prediction, and found it unreliable for the task. As a result we investigated multiple metrics and found the F1@k metric to reflecthuman activity best among them. We demonstrated that our HERA naturally handles hierarchically structured activities, including interruptions in the observed activity hierarchy. When compared to related methods that do not exploit the hierarchical structure in human activities, or explore it in a sub- optimal way, HERA attained superior results specially in the long-term regime.
Acknowledgments We would like to thank Thao Minh Le for the helpful discussions in regards to building the Hierarchical Breakfast dataset.

References

[1] Yazan Abu Farha and Juergen Gall. Uncertainty Aware anticipation of activities. In 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), October 2019.

[2] Icek Ajzen. From intentions to actions: A theory of planned behavior. In Action Control: From Cognition to Behavior. Springer Berlin
Heidelberg, 1985.

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4724–4733, July 2017.

[4] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN Encoder–Decoder for statistical machine translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1724–1734. ACL, 2014.

[5] Junyoung Chung, Sungjin Ahn, and Yoshua Bengio. Hierarchical multiscale recurrent neural networks. In 5th International Conference on Learning

NJICE –National Journal on Information and Communication Engineering
ISSN: 2231-2099 - Volume 6 Issue 1
Apr-Jun 2016 Pages 20-28

Representations, ICLR 2017.
OpenReview.net, April 2017.

[6] Thi Duong, Dinh Phung, Hung Bui, and Svetha Venkatesh. Efficient duration and hierarchical modeling for human activity recognition. Artificial intelligence, May 2009.

[7] Yazan Abu Farha, Alexander Richard, and Juergen Gall. When will you do what? - anticipating temporal occurrences of activities. In 2018
IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 2018.

[8] Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Forecasting future action sequences with neural memory networks. In Proceedings of the British Machine Vision Conference 2019. British Machine Vision Association, September 2019.

[9] Salah El Hihi and Yoshua Bengio. Hierarchical recurrent neural networks for Long-Term dependencies. In Advances in Neural Information Processing Systems 8. MIT Press, 1996.

[10] Qiuhong Ke, Mario Fritz, and Bernt Schiele. Time-Conditioned action anticipation in one shot. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.

[11] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In 2018