PROGRESSIVE KNOWLEDGE DISTILLATION FOR EARLY ACTION RECOGNITION

Vinh Tran, Niranjan Balasubramanian, Minh Hoai

Stony Brook University, Stony Brook, NY 11790

ABSTRACT

We present a novel framework to train a recurrent neural network for early recognition of human actions, which is an important but challenging task given the need to recognize an on-going action based on partial observation. Our framework is based on knowledge distillation, where the network for early recognition is viewed as a student model. The student is trained using knowledge distilled from a more knowledgeable teacher model that can peek into the future and incorporate extra observations about the action in consideration. This framework can be used in both supervised and semi-supervised learning settings, being able to utilize both the labeled and unlabeled training data. Experiments on the UCF101, SYSU 3DHOI, and NTU RGB-D datasets show the effectiveness of knowledge distillation for early recognition, including when we only have a small amount of annotated training data.

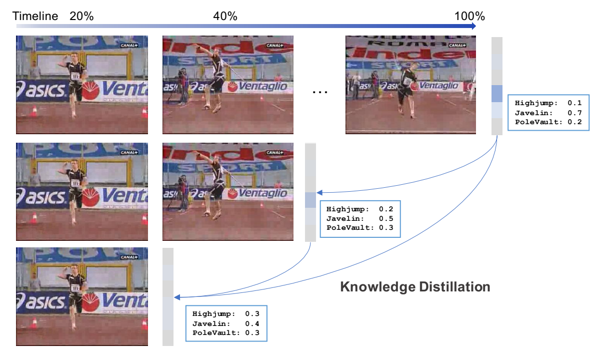


Fig. 1: **Knowledge distillation for early recognition of human actions.** An early classifier can be trained by distilling the knowledge from another or even the same classifier that has privileged access to additional observations about the action in consideration.

1. INTRODUCTION

Early recognition of human action (e.g., [1–12]) refers to the problem of classifying an *ongoing* action, and it is different from the recognition problem, e.g., [13–22]. The former requires classifying partial action sequences, while the latter makes classification decisions based on full observation of the action sequence. Early recognition is crucial in applications that require timely responses, especially for applications in surveillance and human robot interaction.

Training classifiers over partial action sequences is difficult because of the inherent ambiguity in partial action sequences, especially in the early stages of an action where only a small fraction of the action has been performed. Without a proper training procedure, the obtained classifier might not have the *right knowledge* to extract the relevant information about the ongoing action.

In this paper, we propose a novel knowledge distillation framework to train a partial-action classifier, guiding it to attend to the relevant information about the ongoing human action. Under our framework, the partial-action classifier is a recurrent neural network that is trained with distilled knowledge from a teacher network that has superior discriminative power. The teacher is an action recognition network trained on full video sequences or the partial-action classifier itself but with a longer observed action sequence as the input. Our framework is developed based on the intuition that a longer action sequence is less ambiguous than a shorter action sequence, as illustrated in Figure 1, so a network with more observations about the action can act as the teacher. Previous works have attempted to recognize partial actions data using teacher-student framework [19]. However, our framework is more comprehensive with the inclusion of an approach for self-supervised knowledge distillation from a single model.

In our knowledge distillation framework, the target that the student network should output is the probability vector produced by the teacher network, not the binary annotation vector. There are several advantages of using knowledge distillation for early recognition. First, the probability vector produced by the teacher network is a soft target that contains some information about the degree of similarity and correlation between the action categories. This type of information is not encoded in the binary annotation vector. Second, by not defining the training loss on the annotation vector, the student network can be trained without ground truth annotation. Thus, when unlabeled data is available, we can leverage it to improve the performance of the student network. Finally, the soft targets have higher entropy, they contain much more information in a single training sample. As a result, the student network can be trained with much less labeled data.

In summary, the contributions of this paper are three fold. First, we present a general framework for early action

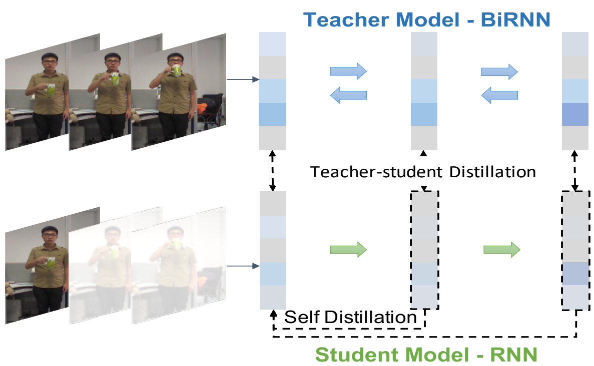


Fig. 2: **Our proposed knowledge distillation framework for early action recognition.**

recognition based on knowledge distillation. Second, we incorporate a novel self-distillation loss into the framework. Finally, we show that the proposed knowledge distillation framework improves the performance of an early recognition network on three human action datasets: UCF101 [23], SYSU 3DHOI [24], and NTU RGB-D [25]. Especially, our proposed method works effectively even with only small amount of labeled training data. With knowledge distillation, we achieve the state-of-the-art early recognition performance on all three datasets.

2. KNOWLEDGE DISTILLATION FRAMEWORK

In this section, we describe the proposed knowledge distillation framework for early recognition of human action.

2.1. Network Architectures

Our framework is based on knowledge distillation, where the desired network for early recognition is the student, and it is trained with the distilled knowledge from a teacher model and also the self-distilled knowledge from the student model when it is allowed to observe more frames, as illustrated in Fig. 2.

We use a one-directional Recurrent Neural Network (RNN) as the student model for early recognition. RNN is particularly suitable for early recognition given its ability to integrate new observations and make predictions at every time step. In particular, we use a one-layer Gated Recurrent Unit (GRU) [26] network as the student model. We do not use a bidirectional GRU (BiGRU), or any bidirectional RNN in general, for early recognition because a bidirectional network is more computationally expensive and cumbersome than an unidirectional network.

Following [19], we use an one-layer BiGRU network as the teacher model. The BiGRU/BiRNN has been widely used for action recognition in videos [19–22]. There are two benefits in using BiGRU as the teacher model. First, it provides a feature representation at each progression level similar to the one directional GRU. Second, since this is bidirectional, the hidden state vector at each time step incorporates both forward and backward information about the action. This hidden vector contains features from both the past and the future at each time step.

2.2. Knowledge Distillation

We use knowledge distillation for training an early recognition network in a novel setting where knowledge distillation does not flow from a complex to a simple model, but from a model with privileged access to more observations to a model with fewer observations. We propose two distillation schemes, one based on the distillation between two separate networks and one based on the self-distillation.

2.2.1. Teacher-Student Distillation.

An input video sequence can be represented as a feature sequence $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N)$ of N progression levels, where $\mathbf{x}_n \in \Re^d$. We use the teacher network $\mathcal{T}$ and the student network $\mathcal{S}$ to compute the prediction outputs at each time step as $\mathcal{T}(\mathbf{x}_n)$ and $\mathcal{S}(\mathbf{x}_n)$, where $\mathcal{T}(\mathbf{x}_n)$ and $\mathcal{S}(\mathbf{x}_n) \in \Re^c$ and c is the number of action classes. Using these prediction outputs, the teacher-student distillation loss for each sequence is then computed as the Kullback–Leibler (KL) divergence between the student and teacher:

$$\mathcal{L}_{ts}(\mathcal{S}, \mathcal{T}) = \frac{1}{N} \sum_{n=1}^{N} KL(\mathcal{T}(\mathbf{x}_n) || \mathcal{S}(\mathbf{x}_n)). \tag{1}$$

Here, we define the knowledge distillation loss based on the KL divergence between two output probability vectors. An alternative approach is to define the distillation loss based on the discrepancy between the two latent representation vectors. However, this requires that the teacher $\mathcal{T}$ and the student $\mathcal{S}$ to have the same latent space, which means we cannot exploit different architectures for $\mathcal{T}$ and $\mathcal{S}$ as we do here.

2.2.2. Self Distillation.

The second distillation scheme comes from the intuition that the recognition accuracy should increase as the ongoing action becomes more complete and the model has more observations about the action. Hence, the recognition output of the model at a later time step can be used as a supervision signal for the recognition output at an earlier time step. We refer to this as the self-distillation loss, which we compute using the KL divergence between the output distributions for a time step n and a later time step $n + \tau$:

$$\mathcal{L}_{self}(\mathcal{S}) = \frac{1}{N - \tau} \sum_{n=1}^{N-\tau} KL(\mathcal{S}(\mathbf{x}_{n+\tau}) || \mathcal{S}(\mathbf{x}_n)), \tag{2}$$

where τ is the lead time for peeking into the future.

2.3. Combined Training Loss

The student model should also output the action category that corresponds to the ground truth label. The loss for the predicted output is defined as:

$$\mathcal{L}_{cls}(\mathcal{S}, y) = \frac{1}{N} \sum_{n=1}^{N} \ell(\mathcal{S}(\mathbf{x}_n), y), \tag{3}$$

where $\ell(\mathcal{S}(\mathbf{x}_n), y)$ is the cross-entropy between the output probabilities $\mathcal{S}(\mathbf{x}_n)$ at time n and the action label y. Finally, the combined loss for training the early recognition network $\mathcal{S}$ defined as:

$$\mathcal{L}(\mathcal{S}, y) = \mathcal{L}_{cls}(\mathcal{S}, y) + \alpha \mathcal{L}_{ts}(\mathcal{S}, \mathcal{T}) + \beta \mathcal{L}_{self}(\mathcal{S}), \tag{4}$$

where α and β are tune-able hyper parameters that control the impact of each knowledge distillation component.

3. EXPERIMENTS

We perform experiments on three datasets and consider both supervised and semi-supervised settings. We compare the proposed method with the direct baseline method that does not use knowledge distillation as well as the other state-of-the-art methods.

3.1. Datasets

We evaluate the proposed knowledge distillation framework for early recognition on three benchmark datasets: UCF101 [23], SYSU 3D Human Object Interaction (SYSU 3DHOI) [24], and NTU RGB-D [25]. Each video is divided into $N = 10$ segments in both training and evaluation. Top-1 accuracy for different observational ratios are reported.

UCF101 dataset comprises of 13,320 action clips from 101 categories collected from YouTube. Following [12, 19], we use the first 15 groups for training, the next three groups for validation, and the rest for testing. Temporal Shift Module (TSM) network [27] model pretrained on the Kinetics [28] dataset is used to extract video features.

SYSU 3DHOI dataset contains 12 activity classes with 480 RGB-D video sequences with 3D skeleton data aptured by a Kinetics camera. We use both RGB (TSM) [27] and skeleton (VA-CNN) [29] features.

NTU RGB-D dataset contains 60 human activities with 56,000 skeleton sequences performed by 40 subjects. We follow [12, 25] and perform experiments on cross-subject settings. VA-CNN [29] is also used to extract skeleton features.

Implementation details. For our early recognition model, we use a one-layer one-directional GRU [26] with hidden size 512 to recognize the action at each time step. The teacher model is a one-layer BiGRU of size 256 in each direction and is trained on fully observed sequences. All models are optimized with SGD of learning rate 0.01. We set $\alpha = 0.5$, $\beta = 0.5$ for UCF101 and $\beta = 1.0$ for SYSU 3DHOI and NTU RGB-D datasets.

	Observational ratio					
	20%	40%	60%	80%	100%	AUC
On the SYSU 3DHOI dataset						
Without distillation	65.4	76.7	81.7	84.2	85.0	76.5
With distillation	**67.1**	**79.2**	**84.2**	**85.8**	**87.1**	**78.8**
On the UCF101 dataset						
Without distillation	90.1	**92.0**	92.6	92.9	93.1	91.7
With distillation	**90.5**	**92.0**	**92.9**	**93.3**	**93.5**	**92.0**

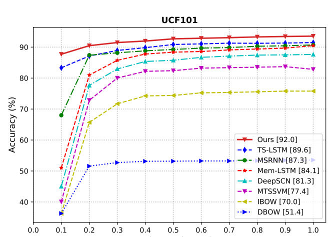
Table 1: **The benefits of knowledge distillation** for early recognition on the SYSU 3DHOI and UCF datasets.

	Observational ratio					
	20%	40%	60%	80%	100%	AUC
RankLSTM [8]	16.5	37.7	55.9	64.4	66.0	43.1
DeepSCN [11]	21.5	39.9	54.6	60.2	58.6	43.2
KNN [12]	9.6	16.0	26.0	34.5	37.0	21.9
MSRNN [12]	20.3	41.4	59.2	67.4	69.2	46.6
TS-LSTM* [19]	22.8	55.3	76.2	85.6	87.8	61.8
Ours	**24.6**	**57.7**	**76.9**	**85.7**	**88.1**	**62.8**

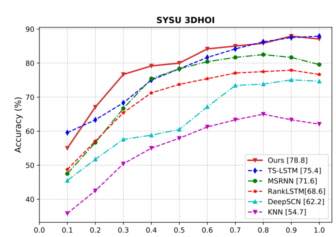
Table 2: **Results on NTU RGB-D dataset.**

3.2. The benefits of knowledge distillation

We first evaluate the benefits of knowledge distillation on the SYSU 3DHOI and UCF101 datasets. We compare the models trained with and without knowledge distillation. As can be seen from Tab. 1, training an early recognition model with knowledge distillation improves the early recognition performance at every observation ratio. The overall early recognition performance AUC for both datasets are also improved, from 91.7% to 92.0% on the UCF dataset and from 76.5% to 78.8% on the SYSU 3DHOI dataset. We also find that both types of knowledge distillation provide benefits. Without the self-distillation loss, the early recognition AUC on the SYSU 3DHOI and UCF datasets are 77.6% and 91.8%, respectively.



(a) UCF101



(b) SYSU 3DHOI

Fig. 3: **Results on UCF101 and SYSU 3DHOI dataset.**

	Training data	Observational ratio 20%	40%	60%	80%	100%	AUC
On the SYSU 3DHOI dataset							
Baseline (w/o knowledge distillation)	10% labeled	43.3	54.6	61.3	63.3	60.0	54.8
TS-LSTM*	10% labeled, 90% unlabeled	41.7	55.0	61.3	60.8	57.9	54.3
Ours	10% labeled, 90% unlabeled	**50.0**	**59.6**	**66.7**	**69.2**	**73.8**	**61.0**
On the UCF101 dataset							
Baseline (w/o knowledge distillation)	10% labeled	83.5	84.8	85.6	86.2	85.6	85.0
TS-LSTM*	10% labeled, 90% unlabeled	84.3	86.4	87.0	87.5	87.5	86.1
Ours	10% labeled, 90% unlabeled	**86.6**	**88.6**	**89.6**	**90.4**	**91.2**	**88.8**

Table 3: **Results on the UCF101 and SYSU 3DHOI datasets with limited amount of labeled training data.** We assume only 10% of the training data is labeled, while the majority 90% of the data is unlabeled. Baseline is the method that only uses classification loss, it does not use knowledge distillation and it cannot utilizes unlabeled data.

3.3. Comparison to the state-of-the-art methods

We also compare our method to the recent state-of-the-art methods on the UCF101 and SYSU3 DHOI datasets. The comparison results are shown in Fig. 3. The proposed method outperforms the other methods significantly on the UCF101 dataset. We improve the state-of-the-art AUC by 2.4% (89.6% $\rightarrow$ 92.0%). The trend is similar on the SYSU 3DHOI dataset. Considering the area under the performance curve (AUC), the proposed method outperforms the other methods by a wide margin. The AUC of the proposed method is 78.8%, which is significantly higher than 75.4% AUC of the second best method TS-LSTM. The performance gains are higher for the smaller observation ratios.

Finally, we compare the proposed method with the state-of-the-art methods on NTU RGB-D dataset. Our model significantly improves the prediction performance on this dataset. The full results are shown in Table 2. Overall, considering the AUC, our method still outperforms TS-LSTM even though TS-LSTM has privileged access to RGB-D features. Our method achieves the new state-of-the-art AUC result of 62.8% on the NTU RGB-D dataset.

3.4. Knowledge distillation with unlabeled data

As mentioned earlier, one benefit of our framework is the ability to leverage unlabeled data. In this experiment, we evaluate the early recognition performance under a semi-supervised learning setting. For this experiment, we pretend that only 10% of the training data comes with annotation, while the majority 90% of the training data is unlabeled. On the labeled portion we can compute both the prediction and distillation losses, while on the portion where the labels are removed, we only compute distillation losses. In this setup, we lower the contribution of the prediction loss $\mathcal{L}_{cls}(\mathcal{S}, y)$ so that we can investigate the effectiveness of the distillation losses during training. We compare the proposed method with the direct baseline method where knowledge distillation is not used and also TS-LSTM*, our reimplementation of TS-LSTM [19] using the feature representation and experimental setup as our method. It can be seen from Tab. 3, the proposed method performs early recognition effectively even with a small amount of labeled training data. On the UCF101 dataset with TSM [27] features, our method has a 3.8% improvement over the direct baseline without distillation and is about 2–3% better than TS-LSTM* at all observational ratios. Similarly, we also observe improvements at all observational ratios in the SYSU 3DHOI datasets. The proposed method achieves the best AUC in both datasets.

4. CONCLUSIONS

We have introduced a framework to improve the training of an early action recognition system using two types of knowledge distillation. The first type of knowledge distillation comes from an external teacher, a bidirectional recurrent neural network with access to the future. The second one is achieved by progressively transferring the knowledge from the same network but with longer observation input sequences. The proposed knowledge distillation framework improves the performance of the early recognition network.

Acknowledgement. This material is based on research sponsored by the Air Force Research Laboratory (AFRL), DARPA, under agreement number FA8750-19-2-1003. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the AFRL, DARPA, or the U.S. Government.

References

[1] Minh Hoai and Fernando De la Torre, "Max-margin early event detectors," *IJCV*, vol. 107, no. 2, pp. 191–202, 2014.

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

[3] Boyu Wang and Minh Hoai, "Predicting body movement and recognizing actions: an integrated framework for mutual benefits," in *Proc. Int. Conf. Autom. Face and Gesture Recog.*, 2018.

[4] Boyu Wang, Lihan Huang, and Minh Hoai, "Active vision for early recognition of human actions," in *Proc. CVPR*, 2020.

[5] Boyu Wang and Minh Hoai, "Back to the beginning: Starting point detection for early recognition of ongoing human actions," in *CVIU*, 2018, vol. 175, pp. 24–31.

[6] Kang Li and Yun Fu, "Prediction of human activity by discovering temporal sequence patterns," *IEEE PAMI*, vol. 36, no. 8, pp. 1644–1657, 2014.

[7] Jun Liu, Amir Shahroudy, Gang Wang, Ling-Yu Duan, and Alex Kot Chichung, "Skeleton-based online action prediction using scale selection network," *IEEE PAMI*, 2019.

[8] Shugao Ma, Leonid Sigal, and Stan Sclaroff, "Learning activity progression in lstms for activity detection and early detection," in *Proc. CVPR*, 2016.

[9] M. Pei, Y. Jia, and S.-C. Zhu, "Parsing video events with goal inference and intent prediction," in *Proc. ICCV*, 2011.

[10] Mohammad Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Basura Fernando, Lars Petersson, and Lars Andersson, "Encouraging lstms to anticipate actions very early," in *Proc. ICCV*, 2017.

[11] Yu Kong, Zhiqiang Tao, and Yun Fu, "Deep sequential context networks for action prediction," in *Proc. CVPR*, 2017.

[12] Jian-Fang Hu, Wei-Shi Zheng, Lianyang Ma, Gang Wang, Jianhuang Lai, and Jianguo Zhang, "Early action prediction by soft regression," *IEEE PAMI*, vol. 41, no. 11, pp. 2568–2583, 2018.

[13] Minh Hoai and Andrew Zisserman, "Thread-safe: Towards recognizing human actions across shot boundaries," in *Proc. ACCV*, 2014.

[14] Yang Wang and Minh Hoai, "Improving human action recognition by non-action classification," in *Proc. CVPR*, 2016.

[15] Yang Wang and Minh Hoai, "Pulling actions out of context: Explicit separation for effective combination," in *Proc. CVPR*, 2018.

[16] Yang Wang, Vinh Tran, and Minh Hoai, "Eigen-evolution dense trajectory descriptors," in *Proc. Int. Conf. Autom. Face and Gesture Recog.*, 2018.

[17] Yang Wang, Vinh Tran, Gedas Bertasius, Lorenzo Torresani, and Minh Hoai, "Attentive action and context factorization," in *Proc. BMVC.*, 2020.

[18] Yang Wang, Gedas Bertasius, Tae-Hyun Oh, Abhinav Gupta, Minh Hoai, and Lorenzo Torresani, "Supervoxel attention graphs for long-range video modeling," in *Proc. WACV*, 2021.

[19] Xionghui Wang, Jian-Fang Hu, Jian-Huang Lai, Jianguo Zhang, and Wei-Shi Zheng, "Progressive teacher-student learning for early action prediction," in *Proc. CVPR*, 2019.

[20] Chunyu Xie, Ce Li, Baochang Zhang, Chen Chen, Jungong Han, Changqing Zou, and Jianzhuang Liu, "Memory attention networks for skeleton-based action recognition," in *Proc. IJCAI*, 2018.

[21] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," in *Proc. ECCV*, 2016.

[22] Jun Liu, Gang Wang, Ping Hu, Ling-Yu Duan, and Alex C Kot, "Global context-aware attention lstm networks for 3d action recognition," in *Proc. CVPR*, 2017.

[23] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah, "UCF101: A dataset of 101 human action classes from videos in the wild," Tech. Rep. CRCV-TR-12-01, University of Central Florida, 2012.

[24] Jian-Fang Hu, Wei-Shi Zheng, Jianhuang Lai, and Jianguo Zhang, "Jointly learning heterogeneous features for rgb-d activity recognition," *IEEE PAMI*, vol. 39, no. 11, pp. 2186–2200, 2017.

[25] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *Proc. CVPR*, 2016.

[26] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. EMNLP*, 2014.

[27] Ji Lin, Chuang Gan, and Song Han, "Tsm: Temporal shift module for efficient video understanding," in *Proc. ICCV*, 2019.

[28] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman, "The kinetics human action video dataset," arXiv:1705.06950, 2017.

[29] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng, "View adaptive neural networks for high performance skeleton-based human action recognition," *IEEE PAMI*, 2019.