Semi-Supervised Few-Shot Atomic Action Recognition

Sizhe Song*
HKUST

Xiaoyuan Ni*
HKUST

Yu-Wing Tai
Kuaishou

Chi-Keung Tang
HKUST

Abstract

Despite excellent progress has been made, the performance on action recognition still heavily relies on specific datasets, which are difficult to extend new action classes due to labor-intensive labeling. Moreover, the high diversity in spatio-temporal appearance requires robust and representative action feature aggregation and attention. To address the above issues, we focus on atomic actions and propose a novel model for semi-supervised few-shot atomic action recognition. Our model features unsupervised and contrastive video embedding, loose action alignment, multi-head feature comparison, and attention-based aggregation, together of which enables action recognition with only a few training examples through extracting more representative features and allowing flexibility in spatial and temporal alignment and variations in the action. Experiments show that our model can attain high accuracy on representative atomic action datasets outperforming their respective state-of-the-art classification accuracy in full supervision setting.

1. Introduction

Significant achievements have been made in action recognition [11, 18, 17] thanks to the development of 3D-convolutional networks (C3D) and spatio-temporal attention over videos. Recent sequential embedding networks including LSTM [10] and temporal convolution (TCN) [12] have been applied for achieving better temporal alignment of a video action. However, the performance of state-of-the-art action recognition models rely on large-scale training datasets which are not easy to collect and annotate. In particular, the pertinent action may not occupy the entire spatio-temporal volume of the given video, i.e., the action may occupy a subset of spatial and temporal volume of the given video frames with intra-class variations in relative position and length. To further complicate the problem, the relative sequences of sub-actions may vary, i.e., 'playing basketball' may contain a different permutation of 'dribbling' and 'passing', which poses great challenges in temporal alignment.

Current methods either ignore alignment such as permutation-invariant attention [23], or impose overly strict alignment such as dynamic time warp [2]. The flexibility within action also presents difficulty in the aggregation of action features within a class. Naïve aggregation functions such as summation may harm representation which may also be easily ruined by outliers.

To tackle the above issues, this paper focuses on *atomic* or fine-grained actions of duration typically less than 2 secs (e.g., dribbling and passing), which sidestep the need for strict alignment while making loose alignment sufficient. Atomic actions have shown promising performance gain for action recognition over conventional methods trained using coarse-grained action videos (e.g., playing basketball) [5]. We propose a novel semi-supervised network for the few-shot atomic action classification, that supports action recognition of long query videos under the K-way N-shot setting [7]. Specifically, our model features better understanding of human-centric atomic action with:

1. *Semi-supervised training.* The video embedding module is trained in a unsupervised manner, extracting representative video features and classifying the action given only several examples.
2. *Loose action alignment.* We adopt sliding windows over the temporal domain and use connectionist temporal classification (CTC) loss [8] to train the video embedding with relatively loose alignment, making the model more robust to variations in sub-action permutation.
3. *Multi-head video comparison.* We develop a multi-head relational network to consider both global and local similarity.
4. *Recursive feature aggregation.* Our model aggregates class features through computing the mutual relationship between support and query videos, and recursively refine the corresponding features. This recursive aggregation extracts the important features for classification, thus reducing the representation complexity for class features and improving classification efficiency.

Overall, this paper contributes to few-shot atomic action recognition with semi-supervised learning. Extensive

*Equal contribution.

experiments over published dataset shows that our method outperforms the state-of-the-art accuracy achieved by models trained in full supervision.

2. Related Work

2.1. Few Shot Learning

Few Shot Learning (FSL) refers to a machine learning strategy that learns from small amount of labeled data where the data labeling cost on large datasets can be prohibitively high [20]. Wide use of FSL includes multiple object tracking and detection [22, 6] and gesture recognition [14, 4]. In this paper, we propose FSL with novel technical contributions in embedding complex human actions through extending the original relational network [15] into a multi-head relation network for robust feature comparison that adequately takes into consideration the high variety of atomic actions, while not requiring a large amount of human-annotated data.

2.2. Action Recognition

Significant progress has been made in action recognition and classification with the recent development of 3D convolutional models, e.g., I3D [3], C3D [19] and TRN [24]. All these models perform data-driven classification and process fixed-size videos by combining the locality in either temporal or spatial domain. Their high accuracy is highly dependent dependency on datasets used in training and testing.

To address the data issue, we experimented a number of FSL for action recognition, and found that almost all of these works attempt align the video in temporal domain and matching the relative 2D frames rather than 3D videos [1, 2], or search for the temporal attention of video [23]. While achieving temporal alignment and attention techniques, these methods partition the given video into individual frames or tiny clips, thus introducing great complexity in their alignment strategies and inevitably losing the generality over datasets with distinct spatio-temporal features. In contrast, our method provides a simple model with holistic understanding over the entire video, focusing on the human-centric video prediction without relying any background and object information.

2.3. Semi-Supervised Learning

Semi-supervised learning is the learning based on both labeled and unlabeled data. In our task, although all the videos all have action labels, there are no boxes to localize where the actions are taking place in individual frames. Thus it is possible to divide the learning strategies into two stages: the first stage is action classification with supervised learning and the second stage is action localization in terms of spatial attention with unsupervised learning. In [23], spatial attention and temporal attention are trained with unsupervised learning.

Typical issues in applying unsupervised learning in feature extraction include limited dictionary size and inconsistent memory. Most recently, the Momentum Contrast (MoCo) has been proposed for unsupervised visual representation learning [9], which regards contrastive learning as dictionary-lookup and builds a dynamic and consistent dictionary on-the-fly. In this paper, MoCo is adopted to pretrain our encoder under an unsupervised setting.

3. Method

Figure 1 illustrates our model structure. We use a C3D+TCN encoder to embed the a video clip to obtain the input feature. The C3D extracts spatio-temporal features from videos and TCN processes the temporal information in a large scale. Next, we apply an attention-pooling module where the support features are refined and integrated. With the query features and refined support features of each class, we then compute the classification probability by a multi-head relation network. Finally, the probability vector and ground truth label are used to obtain a CTC loss and MSE loss to update the network.

3.1. Action Augmentation

We apply the following three augmentation methods:

1. process the videos to produce *human-centric* video tubes, expecting the encoder can thus learn to focus on human bodies in the videos.
2. apply usual image processing such as flipping, blurring, color inverting, rotation.
3. apply background subtraction where the moving object are regarded as foreground and the rest is simply discarded.

The last two methods are easy to implement. For the first method, see Figure 2 where we use a human tracking tool, such as real-time MOT [21] to obtain human-centric video tubes, which are frame-wise bounding boxes of human body with individual's identity labeled in each box. Given these video human tubes, we crop the original frames to get normalized images which precisely capture the movement of each individual in the video. While real-time MOT can generate all the normalized frames on-the-fly, in our modular implementation we generate all the tubes in advance.

3.2. Encoder Pretraining

We use MoCo to pretrain our C3D+TCN encoder. The original video frames are input as support and augmented video frames as query. Then MoCo trains the encoder by a contrastive loss [9] by comparing the query with multiple

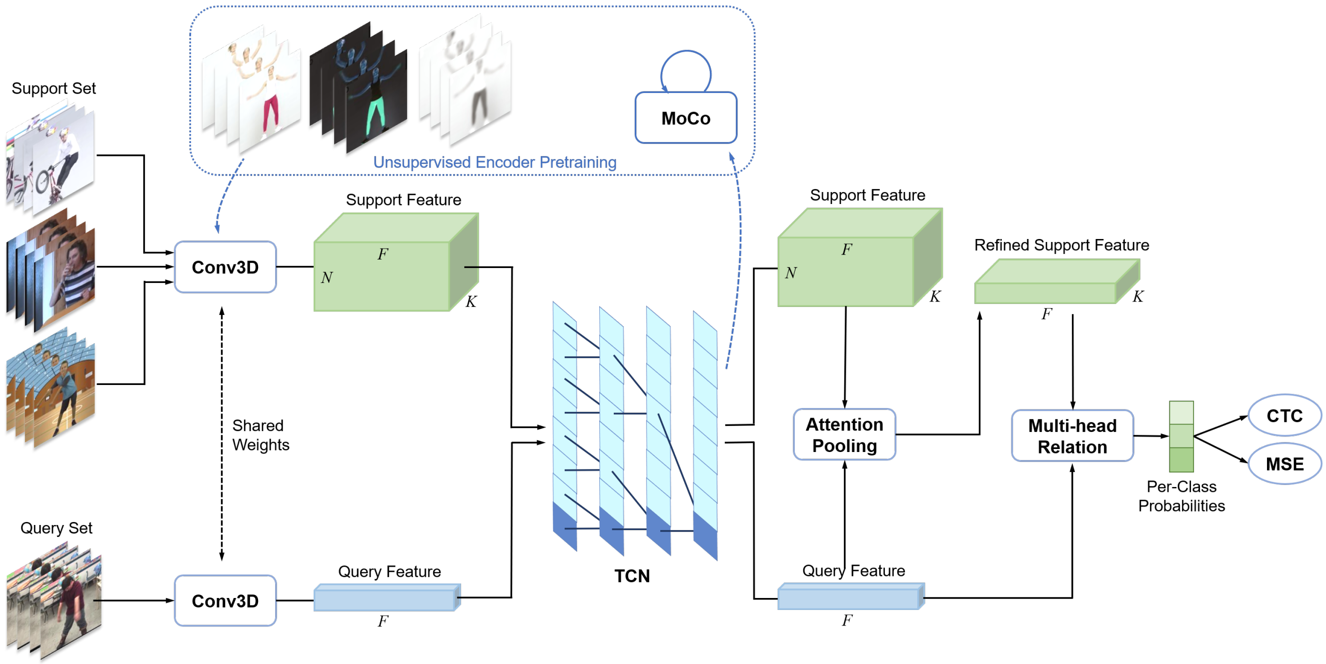

Figure 1. **Overview.** Our model provides fine-grained spatial and temporal video processing with high length flexibility, which embeds the video feature and temporally combines the features with TCN. Further, our models provides attention pooling and compares the multi-head relation. Finally, the CTC and MSE loss enables our model for time-invariant few shot classification training.

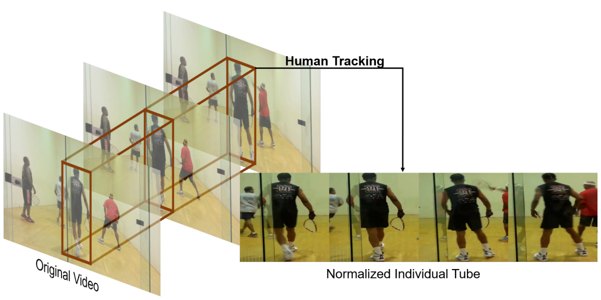

Figure 2. Human-gentric video tube generation.

supports, both positive and negative. This enables the encoder to recognize and extract robust key features from the videos.

MoCo updates the encoder in a momentum way:

$$\theta_k = m\theta_k + (1-m)\theta_q \tag{1}$$

where m is the momentum coefficient, θ_k and θ_q are parameters of the key encoder and query encoder. During back propagation, only θ_q is updated, and then θ_k is updated using this momentum mechanism. This design enables the model to build a consistent memory bank of recently seen data.

3.3. Variable length video processing

For a video V with undefined length, we utilize sliding windows to transfer the video V to a set of windows of fixed-length $\{W_1, W_2, \cdots, W_n\}$. Afterward, each query window feature W_k^q will be compared with a weighted aggregation of support class windows. Then few shot class $S_{k\theta}$ will be represented as:

$$S_{k\theta} = g(W_{ik}^q, A(W_j^s, j \in C_\theta)) \tag{2}$$

where $g(\cdot)$ is the window-wise relational convolution, $A(\cdot)$ is the attention over windows per class which will be detailed in Section 3.4.

After obtaining the class probability for each window, two losses will be computed. The Connectionist temporal classification (CTC) loss is computed for each query video, by aligning all the windows of the query video sequentially. The standard MSE loss will also be computed by adding up the each window's probability for each class:

$$\mathcal{L}_{CTC}(V, l) = \sum_{C:\kappa(C)=l} \prod_{t=1}^{T} P(c_t|W_t) \tag{3}$$

At the test time, beam search decoding is performed to retrieve the video label.

The sliding windows trained with CTC loss can effectively solve the alignment issue for the videos in temporal domain, which is robust against object occlusions, disocclusion and other video instabilities.

3.4. Attention Pooling

Since multiple windows are spread across the support videos to extract support features, a pooling operation is

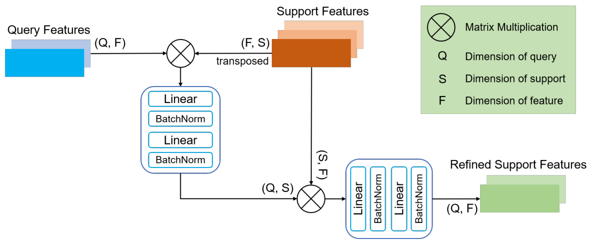

Figure 3. Attention Pooling

necessary to generate fixed-size final representations for relation network where they will be compared with the query.

Commonly used methods are straightforward max-pooling, average-pooling and self-attention. In our work, we propose an attention pooling module. As illustrated in Figure 3, our specially-designed attention pooling takes both the support feature $\mathcal{S} \in \mathbb{R}^{S\times F}$ and query feature $\mathcal{Q} \in \mathbb{R}^{Q\times F}$ as input and computes new support features $\mathcal{S}'$ as follows:

$$\mathcal{S}' = f_2(f_1(\mathcal{Q} \cdot \mathcal{S}^T) \cdot \mathcal{S}) \tag{4}$$

where $f_i(\cdot)$ are linear transformations. The idea is to introduce query features to the pooling procedure by multiplying query features with support features transpose, which will generate a weight matrix $\mathcal{W} \in \mathbb{R}^{Q\times S}$. Each entry $\mathcal{W}_{i,j}$ represents the weight of j^{th} support window to the i^{th} query window. Then the product of this weight matrix and original support features can be seen as the refined support features. Two linear functions f_1 and f_2 are added to provide some learnable parameters.

In addition to support feature refinement, we propose to refine query features by support features. Specifically, the same Equation 4 can be applied except we swap the support and query. This mutual refinement strategy can enhance the prediction performance.

3.5. Multi-Head Relation

The original relation network is composed of a number of convolutional layers [15]. We extend this netowrk to a multi-head relation network in two ways. First, we change the convolution kernel size to 1 because our feature is one-dimensional which no longer retains any spatial structure so the convolution only needs to extract cross-channel information. Second, we add one more computation layer on top of the original Conv+FC layers, which is window-wise vector product. This provides a more localized comparison between the support and query. The final output of multi-head relation network is the sum of the probabilities obtained from both methods.

4. Experiments

4.1. Dataset

We test our model on three datasets including Human-Centric Atomic Action (HAA), Finegym (Gym288) and Moments in Time (MIT). MIT is a general atomic action datasets with over one million videos from diverse classes [13]. Finegym is a recent dataset which focuses on fine-grained gymnastic action classes [16]. HAA provides human-centric videos, with a high average of 69.7% detectable joints [5]. These datasets are not constructed specifically for few-shot learning, so we reorganize them to suit our few-shot task.

HAA & Gym288. Instead of splitting videos into training and test set, we split action classes into training and test set for our few-shot task. Consequently we have 310/156 classes in our HAA training/test set, and 231/57 in our Gym288 respectively.

MIT. The total number of videos in MIT dataset is huge so we build a mini-MIT for our experiment. Each action class in mini-MIT has 60 videos, half from the original training set and another half from the original validation set. Like HAA and Gym288, our mini-MIT has 272/67 classes for training and test respectively.

4.2. Model Performance

Model	**HAA** [5]	**Gym288** [16]	**mini-MIT** [13]
Ours	**80.68**	**75.84**	**61.77**
SOTA	55.33 / 75 top1 / top3	73.7	31.16 / 57.67 top1 / top3

Table 1. Top-1 accuracy (%) on three datasets under a 3-way 5-shot setting compared with state-of-the-art performance in [5, 16, 13] where their models are trained in full supervision.

Table 1 tabulates our results, showing that our few-shot model has a leading performance compared with the state-of-the-art methods on all the three datasets trained in full supervision. Note that our model is few-shot which has access to only very limited amount of data.

5. Conclusion

This paper introduces a novel semi-supervised few-shot action recognition model that outperforms state-of-the-art methods. The sliding window and CTC alignment make the model more robust to coarse temporal annotation. Together with few-shot design, our model can address the intensive data collection and annotation. We also propose the attention pooling and multi-head relation module to achieve better feature refinement and comparison. Further, we incorporate unsupervised learning and different video augmentation methods in video embedding, better enabling the model to recognize decisive features in the videos.

References

[1] Rami Ben-Ari, Mor Shpigel, Ophir Azulai, Udi Barzelay, and Daniel Rotman. Taen: Temporal aware embedding network for few-shot action recognition. *arXiv preprint arXiv:2004.10141*, 2020. 2

[2] Kaidi Cao, Jingwei Ji, Zhangjie Cao, Chien-Yi, Chang Juan, and Carlos Niebles. Few-shot video classification via temporal alignment. In *CVPR*, 2020. 1, 2

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 2

[4] Ka Leong Cheng, Zhaoyang Yang, Qifeng Chen, and Yu-Wing Tai. Fully convolutional networks for continuous sign language recognition. In *ECCV*, 2020. 2

[5] Jihoon Chung, Cheng hsin Wuu, Hsuan ru Yang, Yu-Wing Tai, and Chi-Keung Tang. Haa500: Human-centric atomic action dataset with curated videos. *arXiv preprint arXiv:2009.05224*, 2020. 1, 4

[6] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. In *CVPR*, 2020. 2

[7] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017. 1

[8] Alex Graves, Santiago Fernandez, Faustino Gomez, and Jurgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *ICML*, 2006. 1

[9] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 2

[10] Sepp Hochreiter and Jurgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. 1

[11] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 1

[12] Colin Lea, Rene Vidal, Austin Reiter, and Gregory D. Hager. Temporal convolutional networks: A unified approach to action segmentation. In *ECCV*, 2017. 1

[13] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, and Aude Oliva. Moments in time dataset: one million videos for event understanding. In *IEEE*, 2019. 4

[14] Ruslan Salakhutdinov and Geoffrey Hinton. Deep boltzmann machines. In *International Conference on Artificial Intelligence and Statistics*, 2009. 2

[15] Adam Santoro, David Raposo, David G.T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. *arXiv preprint arXiv:1706.01427*, 2017. 2, 4

[16] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *CVPR*, 2020. 4

[17] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. In *CRCV-TR-12-01*, 2012. 1

[18] Chen Sun, Abhinav Shrivastava, Carl Vondrick, Rahul Sukthankar, Kevin Murphy, and Cordelia Schmid. Relational action forecastingl networks. In *CVPR*, 2019. 1

[19] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 2

[20] Yaqing Wang, Quanming Yao, James Kwok, and Lionel M. Ni. Generalizing from a few examples: A survey on few-shot learning. *arXiv preprint arXiv:1904.05046*, 2020. 2

[21] Zhongdao Wang, Liang Zheng, Yixuan Liu, Yali Li, and Shengjin Wang. Towards real-time multi-object tracking. In *ECCV*, 2020. 2

[22] Xixi Xu, Chao Lu, Liang Zhu, Xiangyang Xue, Guanxian Chen, Qi Guo, Yining Lin, and Zhijian Zhao. Mots: Multiple object tracking for general categories based on few-shot method. *arXiv preprint arXiv:2005.09167*, 2020. 2

[23] Hongguang Zhang, Li Zhang, Xiaojuan Qi, Hongdong Li, Philip H. S. Torr, and Piotr Koniusz. Few-shot action recognition with permutation-invariant attention. In *ECCV*, 2020. 1, 2

[24] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *ECCV*, 2018. 2