MetaVHAR: Meta-Learning for Video-Based Human Activity Recognition

Nasik Muhammad Nafi, William Hsu

Department of Computer Science, Kansas State University
{nnafi, bhsu}@ksu.edu

Abstract

The use and necessity of collaborative robots, particularly personal robots, are increasing at a high rate. Human Activity Recognition (HAR) is an active area of research to improve the performance of personal robots. Activity recognition directly from visual observation is similar in experience with human intelligence. The recent development in deep learning has proven its potential in activity recognition from video data. However, state-of-the-art approaches have validated results on data set that are not suitable for personal robotics applications where proximity and the unique style of the subject are important aspects. In this paper, we present Meta-learning for Video-based Human Activity Recognition (MetaVHAR), a simple but efficient approach for improving generalization to unseen human subjects. We leverage the fact that every human has a unique style of their action and the activity recognition of different humans can be considered as distinct tasks. We validate our approach on a suitable HAR data set UTD-MHAD, which consists of similar actions performed by different humans. Experimental results show that our proposed approach outperforms the baseline classifier trained via standard approach by a large margin.

Introduction

Day by day, personal robots are becoming an integral part of humans to make their life better. They are helping individuals by automating their repetitive and monotonous work at home and office. It's not very far when they will take a similar place like personal computers. Even they are more important to help special groups of people who require more support than others such as the elderly and disabled people. According to the United Nations and the World Health Organization, the number of elderly people is expected to rise by nearly 10% in the upcoming 35 years. As robots need to work alongside different humans, generalization in Human Activity Recognition (HAR) is a crucial aspect to facilitate better human-robot coordination.

The temporal and depth component of human actions make the action recognition task more complex (Tran et al. 2015). The temporal component refers to the features in the temporal dimension. An action can not be identified correctly just from a single frame. The depth channel denotes the distance between the image plane and the corresponding object in the RGB image. RGB-D data has depth information that helps to unambiguously classify the action. Skeleton data, generally referred to as human pose data, captures full information about actions with very limited key points. Wearable inertial sensors such as accelerometers and gyroscopes can provide additional information about an activity. In general, depth information and inertial sensor data are fused to achieve robust performance. However, acquiring all these data has some overhead such as hardware cost, attaching the hardware to the human subject, and extra computational processing compared to the simple video capturing system. Thus, activity recognition based on only RGB video is still at the center of interest.

There has been a lot of progress in the area of deep video classification (Tran et al. 2015) (Carreira and Zisserman 2017) (Arnab et al. 2021). Previous work has attempted to classify a large number of diversified human actions. The most commonly used data sets are sports-1M, Kinetics, UCF-101 (Kay et al. 2017) (Karpathy et al. 2014). These data sets have hundreds of classes including but not limited to smile, talk, run, eat, hair brushing, cycling, snowboarding, dive, ride horse, playing football, fencing, shooting bow, golf, etc. As we can see, the later set of actions is not a good representative of actions happening in a home environment. Also, the angle of view and perspective of the videos are not perfectly aligned with the field of view of a personal robot.

We argue that to improve the performance of a HAR system deployed in a personal robot, the system needs to take into consideration human-specific styles and attributes. At the same time, the system needs to be trained and tested against a benchmark that contains human-specific action data. In this paper, we propose MetaVHAR, Meta-learning for Video-based Human Activity Recognition, which considers the activity recognition of every human as a different task and apply a meta-learning based approach to learning a classifier itself which can generalize better on a new human. We perform our experiment on UTD-MHAD data set consisting of human-specific data. Our experimental results demonstrate that the proposed approach outperforms deep 3d Convolutional Neural Network (3D CNN) based video classifier trained via standard training process across all human subjects in the zero-shot learning scenario.

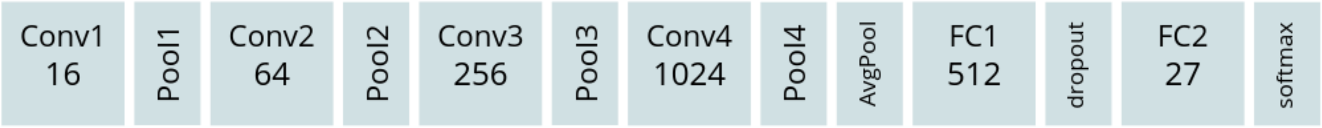


Figure 1: Base Network Architecture

Related Work

Activity Recognition

Activity recognition from the video is a well-studied area of computer vision. While earlier works on video classification have focused on designing hand-crafted spatio-temporal features such as spatio-temporal interest points (STIPs) (Laptev 2005), histogram of oriented gradients(HOG3D)(Scovanner, Ali, and Shah 2007), recent advances in deep learning have triggered the development of deep networks for video. One group of literature has applied CNNs trained on images to extract spatial features from frames followed by a temporal feature processing via Recurrent Neural Network (RNN) or Long Short-Term Memory (LSTM) network (Donahue et al. 2015). Another set of approaches has used 3D CNNs to perform temporal convolutions in video and directly learned the spatio-temporal features (Tran et al. 2015) (Ji et al. 2012). 3D CNNs have achieved higher performance in action recognition tasks when trained on large-scale data sets (Carreira and Zisserman 2017). Recently, transformer and language pretraining-based models have demonstrated superior performance in video classification (Arnab et al. 2021).

Meta-learning

Meta-learning algorithms learn a new task by applying previously acquired knowledge from the set of meta-training tasks. Some approaches consider the meta-learning task as of a recurrent nature where the task-dependent hidden states need to be learned (Santoro et al. 2016)(Munkhdalai and Yu 2017). On the other hand, gradient-based approaches aim to learn the initial parameters of the classifier which can perform better across the tasks (Hochreiter, Younger, and Conwell 2001)(Finn, Abbeel, and Levine 2017). In our work, we adopted a popular gradient-based approach called Model-Agnostic Meta-Learning (MAML) (Finn, Abbeel, and Levine 2017). Recently, meta-learning has been deployed to HAR, however, from a federated learning perspective and using other modalities than video (Li et al. 2021).

Methodology

In this section, we present MetaVHAR - an efficient meta learning-based approach for human activity recognition. We first describe the key idea and then develop the algorithm.

We assume that any human can be considered as a sample drawn from a large distribution of humans, the population. That means all humans share some common features because of their connection to the base distribution while differing from each other. This difference corresponds to the unique style of each human. Thus, the HAR task will be slightly different for each human while sharing the common features. We propose to consider the task of HAR for a particular person τ_i coming from a distribution over tasks $\rho(\tau)$. Hence, $\rho(\tau)$ represents the set of HAR tasks for all humans.

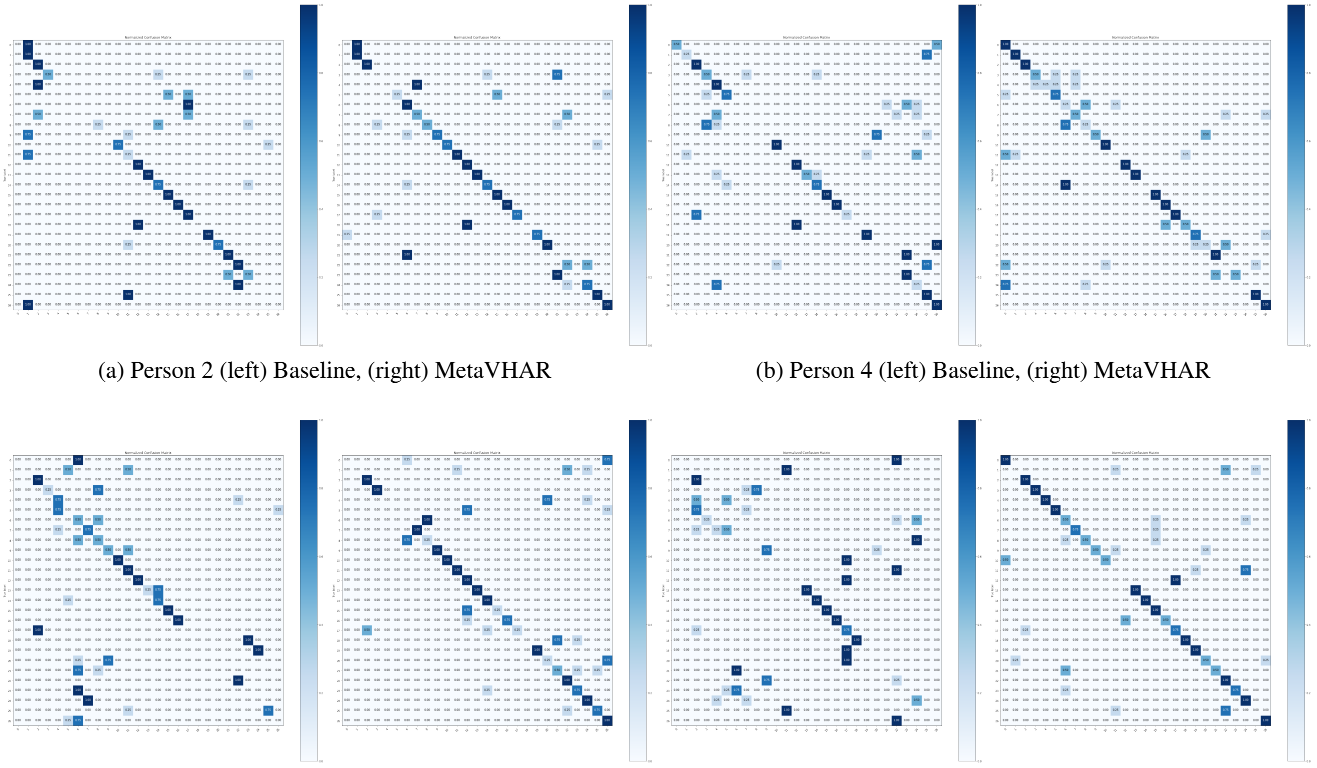
Our objective is to learn the network's initial parameters so that it can perform better on a new task drawn from $\rho(\tau)$. That means we are aiming to learn a classifier network that is suitable for many tasks. Saying another way, we are attempting to learn a HAR system that will produce a good result for a new human sampled from the broader distribution of humans. To do this, we leverage MAML which optimizes the network parameters task-wise so that it can perform better on a new task (Finn, Abbeel, and Levine 2017). Our proposed adaptation of the standard MAML for HAR is described below in pseudocode.

Algorithm 1: MetaVHAR

```
Input: $\tau_{train}$: Sampled humans of dataset for
  training $\tau_{train} \sim \rho(\tau)$
Parameter for Meta-learning: All network
  parameters $\theta$
Output: Learned weights for the network parameters

Initialize network parameters $\theta$ with random weights;
while not done do
    for all human or HAR tasks $\tau_i$ in $\tau_{train}$ do
        1. Sample a batch $D_i = \{x^{(j)}, y^{(j)}\}$ with 1
          data point from each action class for $\tau_i$;
        2. Calculate the loss $\mathcal{L}_{\tau_i}(f_\theta)$ for $D_i$ using $\theta$;
        3. Compute the updated network parameter $\theta'_i$
          with gradient descent:
          $\theta'_i = \theta - \alpha \Delta_\theta \mathcal{L}_{\tau_i}(f_\theta)$;
        4. Sample another batch of samples $D'_i$ for
          that human for the meta-update;
    end
    Update the network parameter $\theta$ with the loss
      calculated from each $\theta'_i$ and $D'_i$:
      $\theta = \theta - \beta \Delta_\theta \sum_{\tau_i \sim \tau_{train}} \mathcal{L}_{\tau_i}(f_{\theta'_i})$;
end
```

We consider the classification network f_θ is parameterized by θ. We attempt to learn more generalizable network parameters starting from an initial set of random weights. Thus the meta-optimization is performed over the network parameters θ. As a proof of concept, we use a simple base network and meta-learned all the network weights via our proposed approach. The network can be considered as a smaller version of the popular C3D model (Tran et al. 2015). Figure 1 presents the base network architecture. We used 4 blocks of 3DConv+Pool followed by a global 3d average



(a) Person 2 (left) Baseline, (right) MetaVHAR

(b) Person 4 (left) Baseline, (right) MetaVHAR

(c) Person 6 (left) Baseline, (right) MetaVHAR

(d) Person 8 (left) Baseline, (right) MetaVHAR

Figure 2: Person-wise confusion matrices for the baseline and proposed method. The presented scores are normalized. The vertical axis denotes the true labels while the horizontal axis denotes the predicted ones.

pooling layer and two fully connected layers. From an experimental perspective, we consider that the whole dataset represents the full human distribution. Thus, τ_{train}, which is a sample subset of $\rho(\tau)$, is considered the training set while the rest of the human in $\rho(\tau)$ is considered as test human.

We update the network in a human-wise (human-specific HAR task-wise) approach. Thus, for each human-specific HAR task τ_i, we sample a batch of data points D_i such that it contains one sample from each action class (line 1 in algorithm). Then the model loss is calculated based on D_i and the current network parameters θ (line 2 in algorithm). As a multi-class classification task, we use the standard cross entropy loss function defined by:

$$\mathcal{L}_{\tau_i}(f_\theta) = \sum_{(x^j,y^j)\sim D_i} \sum_{c=1}^{N} y_c^j \; log f_\theta^c(x^j), \qquad (1)$$

where N is the number of classes, y_c^j is the binary indicator (0 or 1) of whether the class label c is the correct classification for input x^j, and $f_\theta^c(x^j)$ is the corresponding prediction for class c given the input (x^j).

An intermediate set of weights θ_i' is computed with gradient decent using $\theta_i' = \theta - \alpha\Delta_\theta\mathcal{L}_{\tau_i}(f_\theta)$ where α is the step size hyperparameter. Finally, in the outer loop, the meta-objective is calculated across all human-specific HAR tasks in the training set. Particularly, this is estimated based on the loss obtained by equation 1 using the intermediate weights θ_i' and another set of data sampled from that corresponding task. While the network parameter θ is updated through the meta-optimization process, the objective considers the loss evolved from the intermediate weights θ_i'. The human-wise training coupled with the meta-update enables learning the common action-specific features, at the same time, respecting the task-specific (human-specific) differences.

Experiments and Results

Data Set

We use the UTD-MHAD data set which contains 27 actions performed by 8 distinct human subjects (Chen, Jafari, and Kehtarnavaz 2015). There are 4 female and 4 male subjects. Each subject repeated each action 4 times. The data set includes 861 data sequences, missing three of them. As we are formulating different distinct tasks, to have a balance on the data (8 humans × 27 actions × 4 times = 864), we just randomly over-sample three samples from the corresponding action classes for that specific subject. The 27 actions performed are listed in the Appendix. The class list constitutes a comprehensive set of human actions covering hand gestures, daily activities, training exercises, and very few numbers of sports actions. Also, the data is obtained from the close proximity of the subject in an indoor setting.

Table 1: Comparison of MetaVHAR with baseline method on UTD-MHAD Dataset. (*P* refers to ”Person”)

	Accuracy			
Method	**P2**	**P4**	**P6**	**P8**
Baseline	50.93	50.93	44.44	30.56
MetaVHAR	63.89	62.96	62.04	69.44

Experimental Setup

We apply cross subjects protocol to split the training and testing data. As proposed in the original paper, odd human subjects (1,3,5,7) are used for training and even human subjects (2,4,6,8) are used for testing (Chen, Jafari, and Kehtarnavaz 2015). To implement, we use TensorFlow and Keras deep learning library. The baseline classifier has been trained for 200 epochs while the meta-training for 150 epochs. A learning rate of 0.001 and a dropout rate of 0.5 have been used. The size of the input frame is 128 × 128 × 3.

Results and Discussion

We consider the same base 3D CNN-based classifier trained via a standard approach as a baseline. By standard approach we mean, the baseline is trained with all the data from all the training subjects together. We meta-learned the parameters of the same baseline network via MetaVHAR. We used the same learning rate for inner task learning and outer meta-training. Table 1 presents the action recognition accuracy of MetaVHAR compared to the baseline for the 4 subjects in the test set. It is evident that MetaVHAR outperforms the general training-based baseline for all human subjects. The highest performance gain is achieved for human subject 8 which is about 39%. Figure 2 presents the confusion matrices for the baseline and the proposed method for each human subject. The diagonals are more consistent for the MetaVHAR. MetaVHAR enables better representation learning that takes into account the features that are common to all tasks.

Conclusions

We presented a meta learning-based training approach that can efficiently learn to recognize human activity for a new human from videos. Our method consistently demonstrates better performance over the baseline across all unseen human subjects in zero-shot setting. This provides evidence that a meta-learned network for HAR has higher generalization capability than a general classifier. In the future, we plan to meta-learn a state-of-the-art that has already achieved better performance in action recognition. Further, we will extend the work to include other modalities of data that can be obtained directly from videos such as human pose.

References

Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; and Schmid, C. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6836–6846.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.

Chen, C.; Jafari, R.; and Kehtarnavaz, N. 2015. UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *2015 IEEE International conference on image processing (ICIP)*, 168–172. IEEE.

Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; and Darrell, T. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2625–2634.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, 1126–1135.

Hochreiter, S.; Younger, A. S.; and Conwell, P. R. 2001. Learning to learn using gradient descent. In *International Conference on Artificial Neural Networks*, 87–94. Springer.

Ji, S.; Xu, W.; Yang, M.; and Yu, K. 2012. 3D convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1): 221–231.

Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; and Fei-Fei, L. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 1725–1732.

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Laptev, I. 2005. On space-time interest points. *International journal of computer vision*, 64(2-3): 107–123.

Li, C.; Niu, D.; Jiang, B.; Zuo, X.; and Yang, J. 2021. Meta-har: Federated representation learning for human activity recognition. In *Proceedings of the Web Conference 2021*, 912–922.

Munkhdalai, T.; and Yu, H. 2017. Meta networks. In *International Conference on Machine Learning*, 2554–2563.

Santoro, A.; Bartunov, S.; Botvinick, M.; Wierstra, D.; and Lillicrap, T. 2016. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, 1842–1850. PMLR.

Scovanner, P.; Ali, S.; and Shah, M. 2007. A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th ACM international conference on Multimedia*, 357–360.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 4489–4497.

A. Class List

Table 2: List of Classes in the UTD-MHAD Data Set

1	right arm swipe to the left
2	right arm swipe to the right
3	right hand wave
4	two hand front clap
5	right arm throw
6	cross arms in the chest
7	basketball shoot
8	right hand draw x
9	right hand draw circle (clockwise)
10	right hand draw circle (counter clockwise)
11	draw triangle
12	bowling (right hand)
13	front boxing
14	baseball swing from right
15	tennis right hand forehand swing
16	arm curl (two arms)
17	tennis serve
18	two hand push
19	right hand know on door
20	right hand catch an object
21	right hand pick up and throw
22	jogging in place
23	walking in place
24	sit to stand
25	stand to sit
26	forward lunge (left foot forward)
27	squat (two arms stretch out)