CROSS-MODAL KNOWLEDGE DISTILLATION FOR VISION-TO-SENSOR ACTION RECOGNITION

Jianyuan Ni[1] *Raunak Sarbajna*[2] *Yang Liu*[3] *Anne H.H. Ngu*[1] *Yan Yan*[4]

[1] Texas State University, USA
[2] University of Houston, USA
[3] Sun-Yat-Sen University, China
[4] Illinois Institute of Technology, USA

ABSTRACT

Vision modality has been the dominant approach for human activity recognition (HAR), but concerns on how camera systems subject to various viewpoints and occlusions have increased recently. Alternatively, time series data, *i.e.* accelerometer data, from wearable devices can prevent such concerns. However, restricted computational resources associated with wearable devices failed to directly support the advanced deep neural networks with many layers. To tackle this issue and push towards its wide application on HAR understanding, this study introduces an end-to-end Vision-to-Sensor Knowledge Distillation (*VSKD*) framework by transferring the knowledge from vision to sensor domain. To retain the local temporal relationship and facilitate employing visual deep learning models, we convert time series data to two-dimensional images by applying the Gramian Angular Field (*GAF*) based encoding method. We adopted ResNet18 and TSN with BN-Inception as teacher and student network in this study, respectively. After that, we proposed a novel loss function, named Distance and Angle-wised Semantic Knowledge loss (DASK), which is applied to mitigate the intra-modality variations between the video and sensor domain. This study contributes to the field of occlusion-sensitive as well as cross-modal HAR technology. Extensive experimental results on UTD-MHAD, MMAct and Berkeley-MHAD datasets demonstrate the effectiveness and competitiveness of our proposed VSKD model. [Be carefully to talk about privacy issue in vision. Other modalities probably also has privacy issues.]

***Index Terms*—** Cross-modal learning, Knowledge distillation, Human activity recognition, Privacy-sensitive, Signal encoding.

1. INTRODUCTION

Human Activity Recognition (HAR) has been one of the prominent topics, with a focus on perceiving and recognizing actions in various spheres, such as healthcare and human-robot interaction [1]. In recent years, vision-based models have dominated the HAR community due to their popularity and easy access. However, video-based HAR is intrinsically restricted in some occlusion cases and various illumination conditions similar to the human vision limitations. Consequently, such limitations make the video-based approach unfeasible and impractical in such areas. Meanwhile, utilizing time series data, *i.e.* accelerometer data, from wearable devices is another typical way of identifying the HAR problem due to its ability to work in gloomy and bounded conditions [2]. Even though existing methods achieved promising results [3, 4], those methods failed to realize that it is feasible to couple the knowledge from both vision and sensor modalities. For example, vision-based approaches could provide global motion features while sensor-based methods can give 3D information about local body movement [5]. In reality, we understand and perceive the surrounding environment in a multi-modal cognitive way. By utilizing the complementary information acquired from different modalities, we can eventually boost the performance of action recognition. Nevertheless, limited resources associated with the wearable devices, such as CPU and memory storage, cannot support such powerful and advanced multi-modal systems. In order to tackle such issues, the technique of cross-modal transfer, *i.e.* knowledge distillation (KD), is a potential approach that needs only one modality input during the testing phase to reach the performance close to the combination of multi-modal data during the training phase [6]. In this case, we can transfer the knowledge from vision to sensor domain by reducing hardware resource demand from the wearable devices.

[Describe our method first before talking about the contributions.] Based on the above observations, in this study, we propose an end-to-end Vision-to-Sensor Knowledge Distillation (VSKD) to work on the HAR problem. The overview of our proposed method is shown in Figure 1. First, we adopted the Gramian Angular Field (*GAF*) method which encodes the accelerometer data to image representation while keeping the temporal information from accelerometer data [7]. After that, we trained the teacher networks with video stream inputs us-

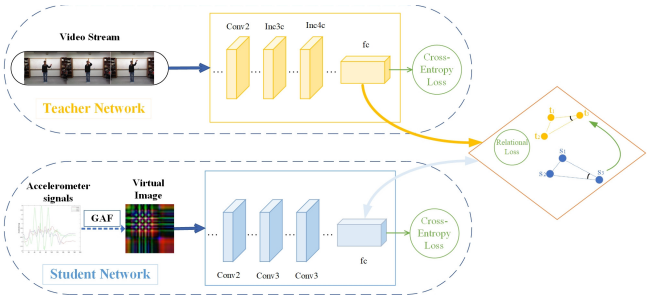

Fig. 1. Schematic overview of our proposed *VSKD* method. [Discuss the figure 1 in the introduction. Also a bit in the caption]

ing cross-entropy loss. The accelerometer data KD process was accomplished by using our proposed loss function. Overall, the contributions of this paper are summarized as follows: 1) To the best of our knowledge, this is the first study conducting the knowledge distillation (KD) model from the video-to-sensor domain. [Are you sure? I think there will be many and you just don't find them] In this *VSKD* model, we use a student network with the input of accelerometer data 2) We proposed a novel loss function, named Distance and Angle-wised Semantic Knowledge loss (*DASK*), which is utilized to alleviate the modality gap between the teacher and student network. Our experimental results confirm the effectiveness of our model and the result on three datasets demonstrate the robustness and competitiveness of our proposed *VSKD* method.

2. RELATED WORK

HAR has been a highly active research field due to its wide applications among various areas, such as healthcare and human-robot interaction [1]. Despite the fact that video modality containing rich RGB information, video modality is often subject to occlusions and various viewpoints or illumination conditions. Moreover, it raises privacy concerns, as videos may capture personal and sensitive information. Consequently, HAR studies with time series data, *i.e.* accelerometer data, from wearable devices have emerged as a promising research field recently [4, 8]. For instance, a wrist-worn tri-axial accelerometer was used to perform arm movement prediction and results demonstrated the robustness of such wearable device [8]. A recurrent neural network (RNN) model was then suggested to deal with such time-dependent input sequences [3]. Additionally, there were some approaches that recommended time series sequences be converted into images in the HAR study [7, 9]. Although those works showed promising results, there is still a significant performance gap between the video and sensor domain on HAR due to intra-modality variations. By aggregating various data modalities, a multi-modal approach can ultimately alleviate the performance gap. For example, Kong *et al.* [10] proposed a multi-modal attention distillation method to model video-based HAR with the instructive side information from time series data. Similarly, Liu *et al.* [9] introduced a multi-modal KD method where the knowledge from multiple sensor data were adaptively transferred to video domain. Although those works provide promising evaluation results on HAR with the multi-modal KD approach, no work has yet been proposed where the sensor domain was applied as the student model within multi-modal KD method. With this framework, it will not only improve the accuracy performance of sensor data on HAR, but also reduce the computational resource demand during the testing phase. Eventually, such framework will be feasible to run on the wearable devices directly.

3. METHODOLOGY

3.1. Virtual Image Generation

[It is better to use bold for vector/matrix in equations.] Inspired by [7], we encodes the accelerometer data to image representation first. In short, we denote one of the three axial accelerometer data (for example, x coordinate) as $\mathbf{X} = \{x_1, x_2, ..., x_n\}$ and normalize it into $\hat{\mathbf{X}}$ among interval [-1, 1]. The normalized $\hat{\mathbf{X}}$ was then encode into the polar coordinate (θ, γ) using the transformation function g. This function encode cosine angle from the normalised amplitude and the radius from the time t, as represented in Eq.1:

$$g\left(\hat{x_i}, t_i\right) = [\theta_i, r_i] \quad \text{where} \quad \begin{cases} \theta_i = arccos(\hat{x_i}), x_i \in \hat{\mathbf{X}} \\ r_i = t_i \end{cases} \tag{1}$$

After this transformation, the correlation coefficient which is equivalent to the cosine of the angle between vectors can be easily calculated upon the trigonometric sum between points [7]. The correlation between time i and j is then calculated using $\cos\left(\theta_i, \theta_j\right)$. Consequently, the tri-axial sensor data with the size of n can be assembled as an image representation $\mathbf{P} = (\mathbf{G_x}, \mathbf{G_y}, \mathbf{G_z})$ of size $n \times n \times 3$. Selected examples of original sensor and GAF-based HAR images of UTD-MHAD [11] are shown in Figure 2.

3.2. DASK Loss

Hinton *et al.* [6] proposed a KD method that compresses knowledge from a larger mode (*i.e. teacher*) into a smaller model (*i.e. student*), while retaining decent accuracy performance. Given a teacher model T_k and a student model S_k, the soft-target $\hat{y}^T$ produced by the teacher model is considered high-level knowledge. The loss of KD when training *student* can be defined as:

$$\mathcal{L}_{KD} = \mathcal{L}_{\mathcal{C}}(y, y^S) + \alpha \mathcal{L}_K(\tilde{y}^T, \tilde{y}^S) \tag{2}$$

$$\mathcal{L}_K = \frac{1}{m} \sum_{k=0}^{m} KL(\frac{P^{T_k}}{T}, \frac{P^{S_k}}{T}) \tag{3}$$

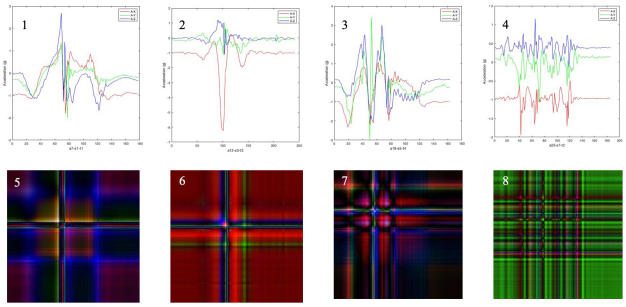

Fig. 2. Original sensor (top) and their corresponding GAF images (bottom) of selected HAR in UTD-MHAD [11] : (1) basketball shooting; (2) bowling; (3) knock on door and (4) walking. [Figure is not clear. Bold and large font for legend and text in the figure please. The same for Fig 1]

where y and y^S refer to the predicted labels and class probability for the student network in this study, respectively. $\tilde{y}^S$ is the "soft target" generated by the student model. Here $\mathcal{L}_C$ is the typical cross-entropy loss and $\mathcal{L}_K$ is the Kullback-Leibler (KL) divergence, while P^{T_k} is the class probability for the teacher network and P^{S_k} is the class probability for the student network. T represents the temperature controlling the distribution of the provability and we use T = 4 in this study according to [6].

However, in order to minimize the intra-modality gap between the vision and sensor domain, we can't just rely on individual predicted outputs themselves. Instead, structural relation and semantic information among those two modalities also needs to be considered [9, 12]. Given a pair of training examples, for instance, the distance-wise function ψ_D tries to minimize the Euclidean distance between teacher and student examples and the distance-wise distillation loss, which tries to penalize the distance differences between teacher and student outputs is defined as:

$$\mathcal{L}_D = \sum_{(x_i,x_j)\in X^2} l_\delta(\psi_D(t_i,t_j),\psi_D(s_i,s_j)) \tag{4}$$

Similarly, the angle-wise distillation loss tries to transfer the relation structures among teacher and students outputs defined as:

$$\mathcal{L}_A = \sum_{(x_i,x_j,x_k)\in X^2} l_\delta(\psi_A(t_i,t_j,t_k),\psi_A(s_i,s_j,s_k)) \tag{5}$$

In addition, since multi-modal data includes the same semantic content, semantic preserving loss is defined as:

$$\mathcal{L}_S = \frac{1}{m}\sum_{k=1}^{m}(\|H^S - H^T\|)_2^2 \tag{6}$$

where H^S and H^T represents the feature of the layer prior to the last fc layer, respectively.

In summary, we use the original KD loss L_{D_K} along with distance and angle-wised distillation loss L_D, L_A, to train the student network and the final DASK loss for the student model is defined as follow:

$$\mathcal{L}_T^S = L_C + \alpha L_K + \beta(L_D + L_A) + \gamma L_S \tag{7}$$

where α, β, γ are the tunable hyperparameters to balance the loss terms for the student network.

4. EXPERIMENTS

4.1. Dataset

In this study, three benchmark datasets were selected due to their multi-modal data forms. We use RGB video streams as the teacher modality and accelerometer data as the student modality in those datasets:

1) *UTD-MHAD* [11]. This dataset covers 27 action classes performed by 8 participants (4 females and 4 males) in quadruplicate. Both modalities have 861 samples and we spit them in half for training and testing.

2) *MMAct* [10]. This dataset includes 37 action classes performed by 20 participants (10 females and 10 males) containing more than 36,000 trimmed clips. Two various settings (cross-subject, and cross-session) are used to evaluate this dataset based on the train and test spit strategy mentioned in [10].

3) *Berkeley-MHAD* [13]. This dataset includes 11 action classes performed by 12 participants (5 females and 7 males) in quintuplicate. In this study, we use the first 7 participants for training and the rest for testing.

4.2. Experimental settings

For the teacher network, we used multi-scale TSN with BN-Inception pre-trained on ImageNet due to its balance between the number of parameters and efficiency [14]. In the teacher network, we set the dropout ratio as 0.5 to reduce the effect of over-fitting. Also, the number of segments is set as 8 for Berkeley-MHAD and UTD-MHAD, while 3 for the MMAct. For the student work, we use ResNet18 as the backbone. All the experiments are running on four Nvidia GeForce GTX 1080 Ti GPUs using PyTorch. We also use a random seed for initializing teacher and student dataloaders to ensure synchronization for both networks. We employed the classification accuracy as the evaluation metric to compare the performance of our VSKD method with other work in which time series data were applied. Also, we adopted F-measure to evaluate the performance of the MMAct dataset according to the previous evaluation strategy [10].

4.3. Experimental results

The comparison results of three datasets are shown in Table 1, Table 2, and Table 3, respectively. In Table 1, our proposed

Method	Testing Modality	Accuracy
Singh et al. [3]	Acc. + Gyro.	91.40
Ahmad and Khan [4]	Acc. + Gyro.	95.80
Wei et al. [5]	Acc. + Gyro.	90.30
Chen et al. [15]	Acc. + Gyro.	96.70
Garcia-Ceja et al. [16]	Acc.	90.20
Our VSKD model	**Acc.**	**96.97**

Table 1. UTD-MHAD Performance Comparison. Accuracy units in %.

Method	Testing Modality	Accuracy
Garcia-Ceja et al. [16]	Acc.	95.40
Mimouna et al. [17]	Acc.	98.0
Das et al. [18]	Acc.	88.90
Our VSKD model	**Acc.**	**99.25**

Table 2. Berkeley-MHAD Performance Comparison. Accuracy units in %.

VSKD model performs better than all the previous comparison models. We make an improvement in testing accuracy of 7.0% compared to the accelerometer view method which extracted 16 features from accelerometer signals for classification [16]. It is worth noting that our proposed VSKD model even performs better as compared to the methods where the accelerometer and gyroscope data were used for testing [3, 4, 5, 15], making an improvement in testing accuracy by 0.5%-6.9%. In Table 2, our proposed VSKD model performs better than all the previous comparison models, increasing the testing accuracy by 1.25 % - 10.35%. Those observations demonstrate that our VSKD method is able to significantly improve the sensor-based HAR problem, because it effectively transfers the information from video to sensor modality. In Table 3, while accelerometer data is the only modality in the testing phase, our method achieves better F-score performance compared to [10, 19] in which either RGB or accelerometer data was used in the testing phase. Similarly, our VSKD approach also outperforms those models in which RGB was applied during the testing phase. This validates that our VSKD approach can effectively learn knowledge from the video domain to improve the accuracy performance of HAR.

4.4. Ablation Study

In order to evaluate the contribution of our proposed DASK loss function, we compare our VSKD with state-of-the-art KD methods [6, 9, 20, 21]. For those KD methods, we use the shared codes, and the parameters are selected according to the default setting. In Table 4, our proposed DASK loss function performs better than all of the previous comparison KD loss functions, indicating that both structural relation and semantic information are critical information in the time series data. Also, angle-wised loss contributes more (0.47 %) to accuracy improvement as compared to distance-wised loss which was consisted with previous work [5], indicating time series data are more valuable to give 3D information about local body movement. Furthermore, compared to structural relation information, semantic information contributes less, which highlights the role of structural relation information on HAR. In addition, even though VGG16 achieves better performance compared to ResNet18 in student baseline, the proposed method with ResNet18 has the best accuracy (96.97%) compared to VGG16 (95.34%).

Method	Testing Modality	Cross Subject	Cross Session
Kong et al. [10]	RGB	66.45	74.58
Kong et al. [10]	wearable	62.67	70.53
Kong et al. [19]	RGB	65.10	62.80
Our VSKD model	**wearable**	**67.83**	**75.72**

Table 3. MMAct Performance Comparison. F-score units in %.

Method	Modality	Accuracy	F1 score
ST [6]	Acc.	96.04	96.15
SP [20]	Acc.	96.50	
CC [21]	Acc.	96.40	
DASK-VGG16	Acc.	95.34	95.69
DASK-ResNet18	Acc.	**96.97**	**96.38**
ASK (W/O D)-ResNet18	Acc.	96.73	96.27
DSK (W/O A)-ResNet18	Acc.	96.51	95.80
SK (W/O D and A)-ResNet18	Acc.	96.50	96.06
DAK (W/O S)-ResNet18	Acc.	95.80	96.04

Table 4. Ablation study of accuracy and F1 score performance (%) on UTD-MHAD dataset. W/O denotes Without.

5. CONCLUSION

In this paper, we propose an end-to-end Vision-to-Sensor Knowledge Distillation (*VSKD*) model to improve the HAR performance. We also propose a novel loss function (DASK), which is able to alleviate the intra-modality gap between vision and sensor modality. Extensive experiments on three multi-modal benchmarks demonstrate the effectiveness and competitiveness of our proposed VSKD method for knowledge transfer from vision to the sensor domain.

6. REFERENCES

[1] Michalis Vrigkas, Christophoros Nikou, and Ioannis A Kakadiaris, "A review of human activity recognition methods," *Frontiers in Robotics and AI*, vol. 2, pp. 28, 2015.

[2] Muhammad Ehatisham-Ul-Haq, Ali Javed, Muhammad Awais Azam, Hafiz MA Malik, Aun Irtaza, Ik Hyun Lee, and Muhammad Tariq Mahmood, "Robust human activity recognition using multimodal feature-level fusion," *IEEE Access*, vol. 7, pp. 60736–60751, 2019.

[3] Satya P Singh, Madan Kumar Sharma, Aimé Lay-Ekuakille, Deepak Gangwar, and Sukrit Gupta, "Deep convlstm with self-attention for human activity decoding using wearable sensors," *IEEE Sensors Journal*, vol. 21, no. 6, pp. 8575–8582, 2020.

[4] Zeeshan Ahmad and Naimul Mefraz Khan, "Multidomain multimodal fusion for human action recognition using inertial sensors," in *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*. IEEE, 2019, pp. 429–434.

[5] Haoran Wei, Roozbeh Jafari, and Nasser Kehtarnavaz, "Fusion of video and inertial sensing for deep learning–based human action recognition," *Sensors*, vol. 19, no. 17, pp. 3680, 2019.

[6] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[7] Feri Setiawan, Bernardo Nugroho Yahya, and Seok-Lyong Lee, "Deep activity recognition on imaging sensor data," *Electronics Letters*, vol. 55, no. 17, pp. 928–931, 2019.

[8] Madhuri Panwar, S Ram Dyuthi, K Chandra Prakash, Dwaipayan Biswas, Amit Acharyya, Koushik Maharatna, Arvind Gautam, and Ganesh R Naik, "Cnn based approach for activity recognition using a wrist-worn accelerometer," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2017, pp. 2438–2441.

[9] Yang Liu, Keze Wang, Guanbin Li, and Liang Lin, "Semantics-aware adaptive knowledge distillation for sensor-to-vision action recognition," *IEEE Transactions on Image Processing*, 2021.

[10] Quan Kong, Ziming Wu, Ziwei Deng, Martin Klinkigt, Bin Tong, and Tomokazu Murakami, "Mmact: A large-scale dataset for cross modal human action understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8658–8667.

[11] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz, "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)*. IEEE, 2015, pp. 168–172.

[12] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho, "Relational knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3967–3976.

[13] Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, and Ruzena Bajcsy, "Berkeley mhad: A comprehensive multimodal human action database," in *2013 IEEE Workshop on Applications of Computer Vision (WACV)*. IEEE, 2013, pp. 53–60.

[14] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba, "Temporal relational reasoning in videos," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 803–818.

[15] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz, "Fusion of depth, skeleton, and inertial data for human action recognition," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 2712–2716.

[16] Enrique Garcia-Ceja, Carlos E Galván-Tejada, and Ramon Brena, "Multi-view stacking for activity recognition with sound and accelerometer data," *Information Fusion*, vol. 40, pp. 45–56, 2018.

[17] Amira Mimouna, Anouar Ben Khalifa, and Najoua Essoukri Ben Amara, "Human action recognition using triaxial accelerometer data: selective approach," in *2018 15th International Multi-Conference on Systems, Signals & Devices (SSD)*. IEEE, 2018, pp. 491–496.

[18] Avigyan Das, Pritam Sil, Pawan Kumar Singh, Vikrant Bhateja, and Ram Sarkar, "Mmhar-ensemnet: A multi-modal human activity recognition model," *IEEE Sensors Journal*, vol. 21, no. 10, pp. 11569–11576, 2020.

[19] Quan Kong, Wenpeng Wei, Ziwei Deng, Tomoaki Yoshinaga, and Tomokazu Murakami, "Cycle-contrast for self-supervised video representation learning," *arXiv preprint arXiv:2010.14810*, 2020.

[20] Frederick Tung and Greg Mori, "Similarity-preserving knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1365–1374.

[21] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang, "Correlation congruence for knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5007–5016.