

Supplementary Material

Multi-Modality Co-Learning for Efficient Skeleton-based Action Recognition

ACM Reference Format:

. 2024. Supplementary Material Multi-Modality Co-Learning for Efficient Skeleton-based Action Recognition. In . ACM, New York, NY, USA, 3 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 OVERVIEW

The supplementary material sections are as follows: Section 2 presents a more detailed description of the datasets and a more detailed introduction to multimodal data processing. Section 3 presents detailed results on ablation studies and zero-shot action recognition to demonstrate the effectiveness and generalization of our proposed MMCL. Section 4 showcases more detailed visualizations to provide a more intuitive demonstration of the modeling capability and robustness of our proposed MMCL.

2 EXPERIMENTAL DATASETS AND DATA PROCESSING

2.1 Experimental datasets

NTU-RGB+D (NTU 60) suggests two benchmark scenarios for evaluation: (1) Cross-View (X-View), where the training data originates from cameras at 0° (view 2) and 45° (view 3), and the testing data is sourced from the camera at 45° (view 1). (2) Cross-Subject (X-Sub), where the training data comprises samples from 20 subjects, while the remaining 20 subjects are reserved for testing.

NTU-RGB+D 120 (NTU 120) suggests two criteria: (1) Cross-subject (X-Sub), where the training data is sourced from 53 subjects, while the testing data originates from the other 53 subjects. (2) Cross-setup (X-Set), where the training data is composed of samples with even setup IDs, and the testing data comprises samples with odd setup IDs.

Northwestern-UCLA (NW-UCLA) adheres to the following evaluation criteria: the training data is derived from two cameras, while the third camera's samples are reserved for testing purposes.

UTD-MHAD consists of 27 actions performed by 8 subjects. Each subject repeated the action for 4 times, resulting in 861 action sequences. The RGB, depth and skeleton were recorded. We employ 191 action samples from the UTD-MHAD dataset for zero-shot action recognition, covering six distinct action categories.

SYSU-Action consists of 12 different actions performed by 40 participants. For each action sample, each participant manipulates

one of the six different objects: phone, chair, bag, wallet, mop and besom. Therefore, there are a total of 480 video clips collected in this set. The contained action samples have different durations, ranging from 1.9s to 21s. We employ 120 action samples from the SYSU-Action dataset for zero-shot action recognition, covering three distinct action categories.

2.2 Multimodal Data Processing

Our MMCL uses four different skeleton modalities, namely joint (J), bone (B), joint motion (JM) and bone motion (BM). Given two joints $v_i = \{x_i, y_i, z_i\}$ and $v_j = \{x_j, y_j, z_j\}$, a bone is defined as a vector $e_{v_i, v_j} = (x_i - x_j, y_i - y_j, z_i - z_j)$. Given two joints data v_{ti} , $v_{(t+1)i}$ and two bones data $e_{v_{(t+1)i}, v_{(t+1)j}}$, $e_{v_{ti}, v_{tj}}$ from two successive frames, the joint motion and bone motion data are defined as $m_{ti} = v_{(t+1)i} - v_{ti}$ and $m_{v_{ti}, v_{tj}} = e_{v_{(t+1)i}, v_{(t+1)j}} - e_{v_{ti}, v_{tj}}$ respectively.

3 DETAILED RESULTS

In Tab. 11, we explored the impact of whether or not using a human detector on the recognition accuracy in MMCL. The results in Tab. 11 indicate that the recognition accuracy can be improved when a detector is used to filter out environmental noise. In our MMCL, we use well-trained weights for zero-shot action recognition in the SYSU-Action and UTD-MHAD datasets. Tab. 12 shows the accuracy of each action category and the overall accuracy in the SYSU-Action dataset, where the data used covers three action categories with a total of 120 samples. Similarly, Tab. 13 illustrates the accuracy for individual categories as well as the overall accuracy of all samples on the UTD-MHAD dataset, where the data used in our experiment comprises six action categories with a total of 191 samples. Compared to the state-of-the-art methods, our MMCL demonstrates a significant improvement in the overall accuracy of all actions and recognition performance of challenging actions (e.g. sitting chair and squat) under the same inference conditions, which indicates the effectiveness and generalization of our MMCL.

4 MORE VISUALIZATION

In Fig. 7, we conducted TSNE visualization of the skeletal features modeled by MMCL. Our observation is two-fold based on the Fig. 7. First, the skeleton feature of the same action category is close and the skeleton feature of different action categories is far apart in the feature space. Second, action samples of different categories but similar are close in the feature space. For example, the 'wield knife' and 'shoot with gun' actions in Fig. 7 are very close in the feature space. Fig. 7 visually demonstrates the skeletal features modeled by our MMCL and provides an interpretation from the feature space of why similar action samples are challenging to differentiate. Meanwhile, the Fig. 8 has showcased more action categories that benefit from our proposed MMCL.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference'17, July 2017, Washington, DC, USA

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/18/06

<https://doi.org/XXXXXXX.XXXXXXX>

Table 11. Comparison of whether to use a detector.

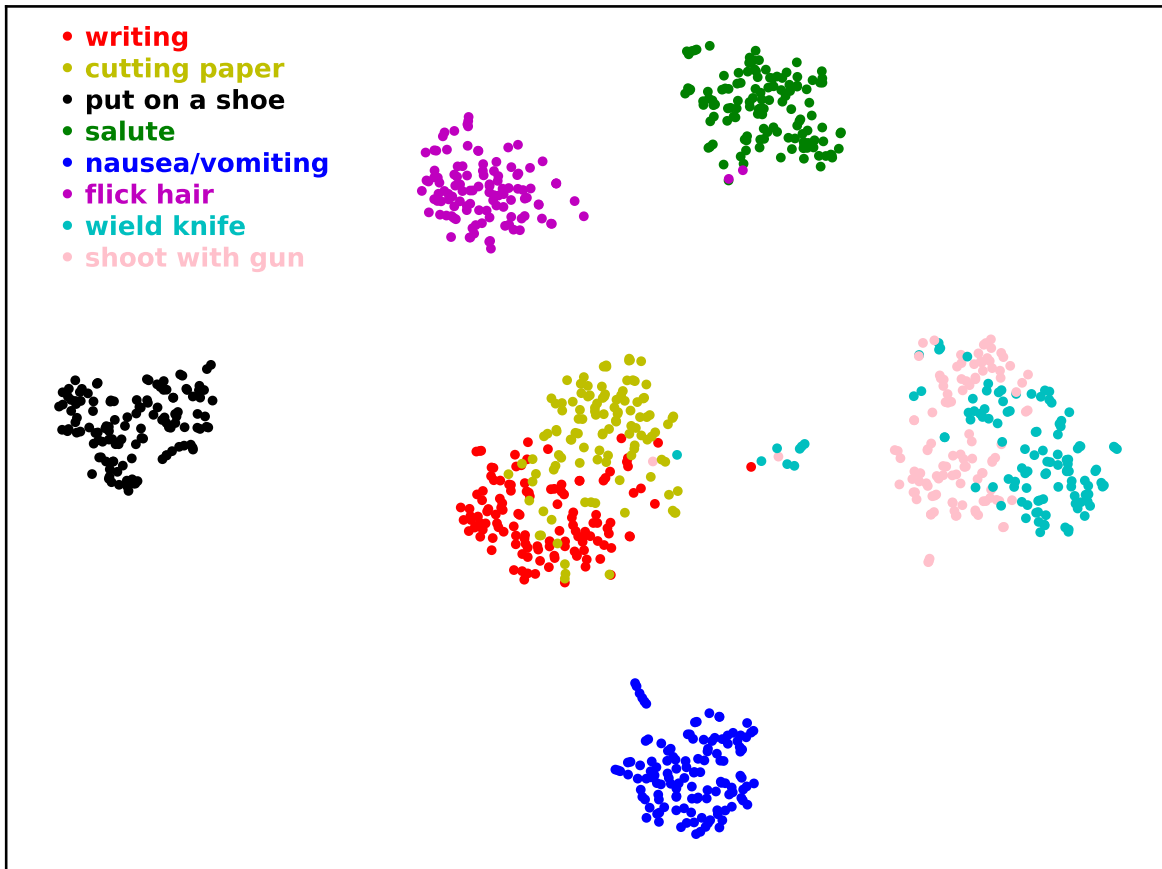
Methods	Detector	Acc. (%)
Baseline	-	85.01
MMCL w/o Detector	-	85.22 ^{↑0.21}
MMCL w/ Detector	YoloV5	85.79^{↑0.78}

Table 12. The accuracy for each category and the overall accuracy in the SYSU-Action dataset.

Methods	Top-1/Top-5 Acc.(%)			
	drinking	calling phone	sitting chair	total
CTRS-GCN (J)	72.50/92.50	40.00/62.50	0.00/5.00	37.50/53.33
CTR-GCN (J)	30.00/55.00	52.50/92.50	0.00/7.50	27.50/51.67
HD-GCN (J CoM-1)	65.00/92.50	15.00/50.00	0.00/32.50	26.27/58.33
Ours (w/o BLIP Refine)	67.50/100.00	45.00/77.50	5.00/52.50	39.17/76.67

Table 13. The accuracy for each category and the overall accuracy in the UTD-MHAD dataset.

Methods	Top-1/Top-5 Acc.(%)						
	wave	arm cross	basketball shoot	sit to stand	stand to sit	squat	total
CTRS-GCN (J)	6.25/37.50	6.25/65.63	34.38/100.00	84.38/100.00	93.75/100.00	0.00/12.90	37.70/69.63
CTR-GCN (J)	3.13/56.25	31.25/65.63	59.38/100.00	100.00/100.00	93.75/100.00	0.00/25.81	48.17/74.87
HD-GCN (J CoM-1)	50.00/87.50	0.00/34.38	28.13/93.75	100.00/100.00	96.88/100.00	0.00/32.26	46.07/74.87
Ours (w/o BLIP Refine)	46.88/93.75	12.50/37.50	46.88/100.00	100.00/100.00	100.00/100.00	9.68/53.13	52.88/81.16

**Figure 7. TSNE visualization regarding the specific action features modeled by our MMCL.**

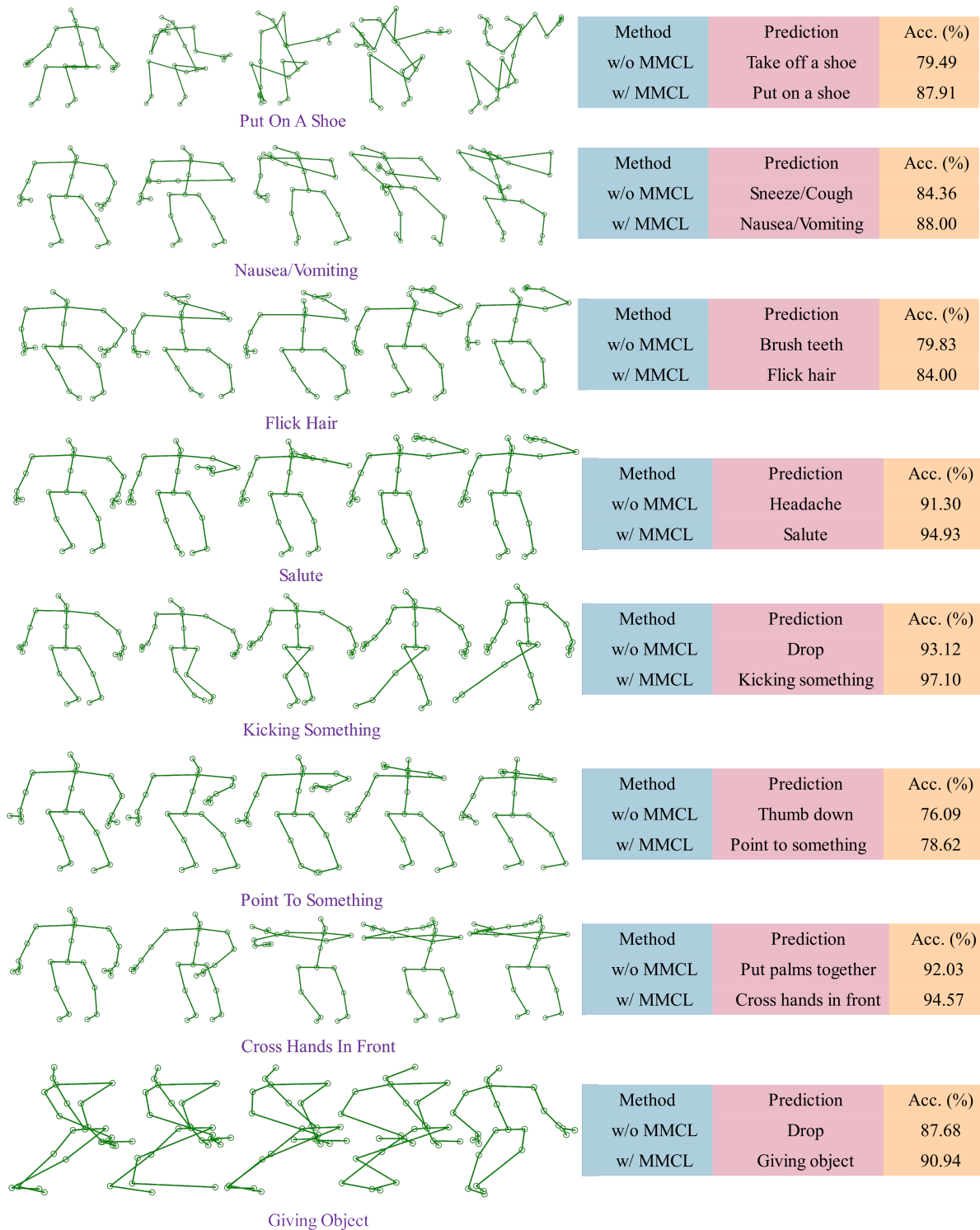


Figure 8. More visualization of improved accuracy about difficult action samples when CTR-GCN used MMCL. The second column represents the prediction of models for the currently visualized sample and the third column represents the accuracy for all samples within the currently visualized category.