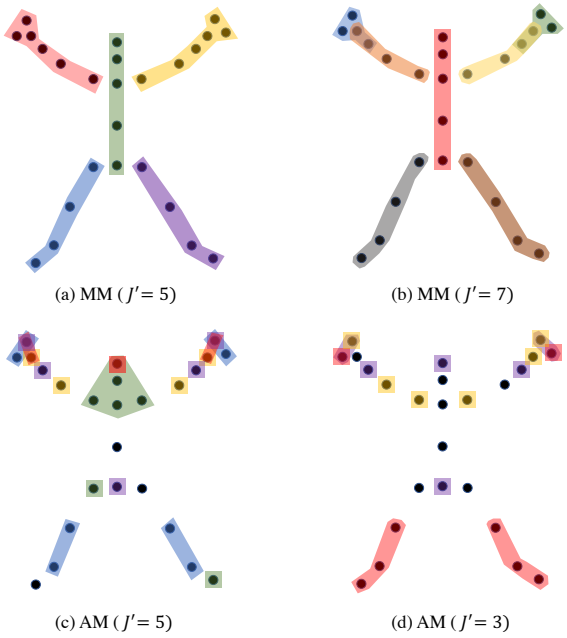
Efficient Skeleton-Based Action Recognition via Joint-Mapping strategies

Min-Seok Kang, Dongoh Kang, HanSaem Kim
Kakao Enterprise
Pangyo, Gyeonggi-do, South Korea
{ahstarwab, kito9021, kensaem}@gmail.com



(a) MM ($J'=5$) (b) MM ($J'=7$) (c) AM ($J'=5$) (d) AM ($J'=3$)

Figure 1. Illustration of the comparison between two manual mappings ($J'=5$ and 7) and adaptive mappings ($J'=5$ and 3).

1. Ablations on J'

We carry out the ablation studies on the different number of decreased joints J' on NTU RGB+D (NTU60) dataset, and the results are reported in Table 1. For MM with $J'=7$ illustrated in Figure 1-(b), additional bottleneck fusion module (*i.e. hand* encoder) is introduced with the observation of action classes that the subtle hand movement is important to distinguish the actions (*e.g. "reading", "writing" and "type on keyboard"*). We also conduct ablation study on varying J' choices for AM module with the observation that there are some redundant mappings for AM with $J'=5$ as shown in Figure 1-(c). For example, fingers correspond to red, blue and purple nodes. On the other hand, we found that the mapping result of AM with $J'=3$ shows more straightforward mapping result, which can be divided as follows: (1.two fingers and body), (2.two legs), and (3.two arms). The corresponding result is illustrated in Figure 1-(d). As shown in Table 1, varying J' choices for the two mapping modules have slight influence over the accuracy, and we set $J'=5$ for both MM and AM as our default, since they show consistent performance on the NTU60.

Type	J'	NTU60 (%)		FLOPs (G)
		X-sub	X-view	
MM	5	90.8	95.2	0.84
MM	7	90.4	95.1	0.85
AM	3	90.6	94.8	0.77
AM	4	90.0	95.0	0.78
AM	5	90.3	95.2	0.78
AM	6	90.2	95.2	0.78
AM	7	90.7	94.9	0.79

Table 1. Ablation studies on J' on the NTU RGB+D (NTU60) dataset.

2. Detail Parameters

We report the number of channels of input and output in Table 2. $input$ and $GC\text{-}TC$ modules in Table 2 are CTR-GC blocks. Specifically, $input$ layers process joint representation on each branch (*joint, bone, and velocity*), while $GC\text{-}TC$ layers process the concatenated representations. $mapping$ is the proposed joint-mapping modules, and $conv$ is the convolutional layer with J'-sized kernels for processing the decreased nodes. We set the size and stride of the kernel at the $conv$ layer to 5 and 3, respectively. fc is a simple fully-connected layer for the final classification.

3. Discussion on Grad-CAM results

We select four different sets of actions depending on the dominant parts for recognizing actions. The first set contains the classes where the movement of arms is critical

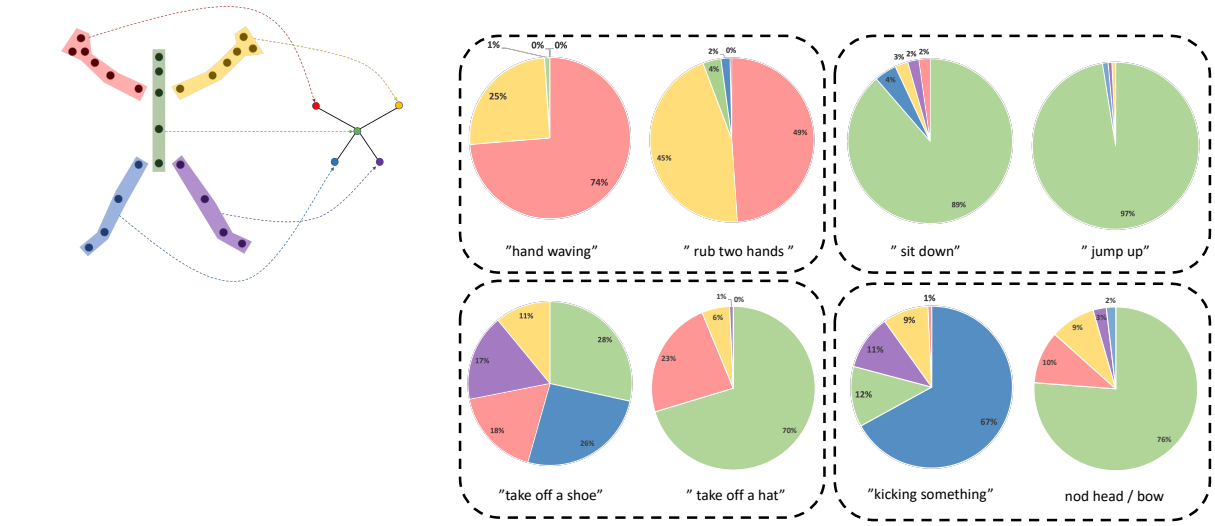
Layer	Main Models		Positions (Table 1)		#Inputs (Table 2)		w/o Map. (Table 3)	
	MM	AM	Pos5	Pos7	J	JB	X-6	X-7
$input_1$	3/64	3/64	3/64	3/64	3/64	3/64	3/64	3/64
$input_2$	64/48	64/48	64/48	64/48	64/64	64/48	64/48	64/48
$input_3$	48/16	48/16	48/16	48/16	64/64	48/24	48/16	48/16
$concat$	16/48	16/48	16/48	16/48	-	24/48	16/48	16/48
$GC\text{-}TC_1$	48/64	48/64	48/64	48/64	64/64	48/64	48/64	48/64
$GC\text{-}TC_2$	64/128	64/128	64/128	64/128	64/128	64/128	64/128	64/128
$GC\text{-}TC_3$	128/128	128/128	-	128/128	128/128	128/128	128/128	128/128
$GC\text{-}TC_4$	-	-	-	128/128	-	-	128/128	128/128
$GC\text{-}TC_5$	-	-	-	128/128	-	-	-	128/128
$mapping_1$	128/128	128/128	128/128	128/128	128/128	128/128	-	-
$conv$	128/128	128/128	128/128	128/128	128/128	128/128	-	-
fc	128/$\#cls$	128/$\#cls$	128/$\#cls$	128/$\#cls$	128/$\#cls$	128/$\#cls$	128/$\#cls$	128/$\#cls$

Table 2. Detailed parameters of the two main models and the models in ablation studies. The reported numbers are the number of channels of input and output on each layer.

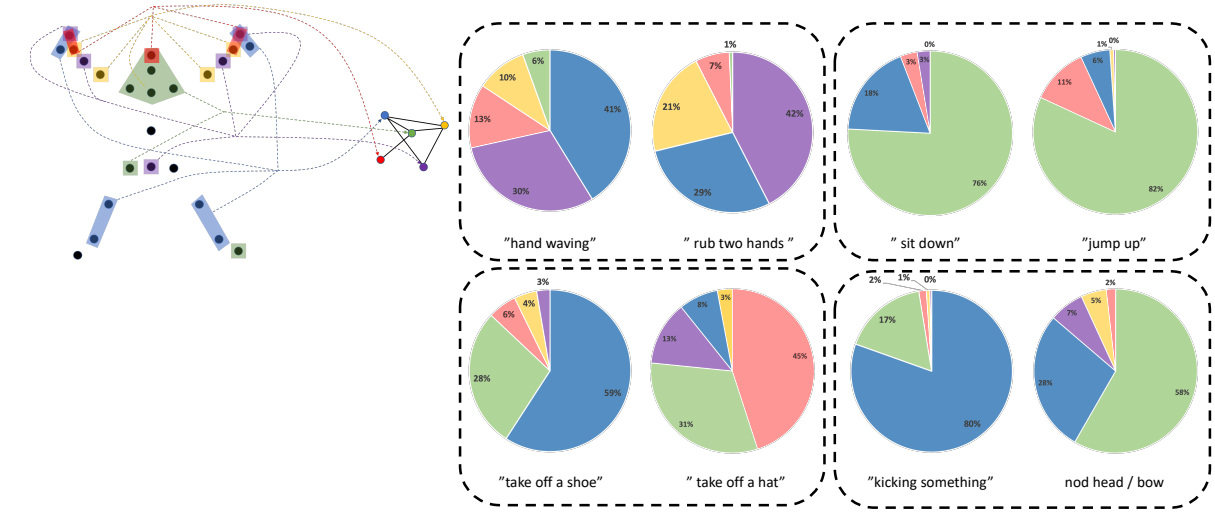
to distinguish the actions (*"hand waving" and "rub two hands"*). The second set includes the classes where the movement of human body is critical (*"sit down" and "jump up"*). The third set includes the classes where the interaction between different parts is critical (*"take off a shoe" and "take off a hat"*). The fourth set contains the action classes where the movement of legs or head contributes the most to distinguish the action classes (*"kicking something" and "nod head / bow"*).

The visualization of MM is illustrated in Figure 2-(a). The mapping result of MM and pie charts indicate quite straightforward results. For distinguishing the first and second set of actions, the arm (red or yellow node) and body (green node) parts contribute the most to the action classification, respectively. As illustrated in the pie charts of the third set, more than two parts contribute to the final classification of *"take off a shoe"*, while the interaction between body (green node) and hand (red or yellow node) is critical to distinguish *"take off a hat"*. For distinguishing *"kicking something"* in the fourth set, the leg part (blue or purple node) is dominant to classify kicking action, but the body part (green node) also contributes to decide the correct action. For distinguishing *"nod head / bow"* in the fourth set, the upper body (green node) is the most critical part for deciding the correct action, but the movement of the two arms (red and yellow node) is also important.

The visualization of AM is illustrated in Figure 2-(b). The mapping result of AM and pie charts are far more complicated compared to the results of MM for the reason that the mapping result of AM is an approximation derived from mapping matrix $\mathbf{M}_G$. That is, all nodes contribute to the final classification regardless of action class. We observe that all the nodes illustrated in blue, purple, red, and yellow contain the joints of two arms. From this observation, we suppose that the movement of two arms is mostly important to classify action classes in NTU RGB+D. We name the five nodes from the approximated mapping result by the joint positions as follows : *blue node (hands and legs)*, *red node (hands and head)*, *green node (upper body)*, *purple and yellow nodes (two arms)*. The upper body (green) barely contributes to classify *"hand waving"* and *"rub two hands"*, while it contributes the most in classifying *"sit down"* and *"jump up"*. When distinguishing *"take off a shoe"*, the blue node including two hands and legs contributes the most to decide the correct action, while the red node including two hands and head is a critical factor in classifying *"take off a hat"*. Through the observation of the pie charts of *"kicking something"* and *"nod head / bow"*, blue and green nodes are quite important to capture the movement of legs and upper body, respectively.



(a) Manual Mapping (MM)



(b) Adaptive Mapping (AM)

Figure 2. Illustration of the mapping results and the pie charts derived from Grad-CAM results across NTU RGB+D dataset. The figures on the left demonstrate what the mapping results of the two joint-mappings look like. Besides, the pie charts on the right indicate which part is more activated for recognizing different action categories. The eight pie charts are categorized into four different sets depending on the dominant parts for recognizing actions. (Best viewed in color and zoomed images).