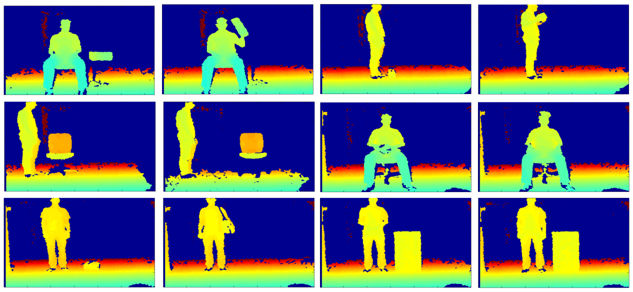
3D Action Recognition from Novel Viewpoints
(Supplementary Material)

Comparison of fc_6 and fc_7 Layer Features

Feature	HPM		HPM+TM	
	N-UCLA	UWA3DII	N-UCLA	UWA3DII
fc_6	76.9	58.5	89.6	76.4
fc_7	**78.1**	**58.9**	**92.0**	**76.9**

Table 1: This table compares the recognition accuracies of fc_6 and fc_7 layer features of our proposed CNN model on the two multiview datasets. Although performance is not very sensitive to these two layers, we use the outputs of the fc_7 layer as the frame descriptors in all experiments.

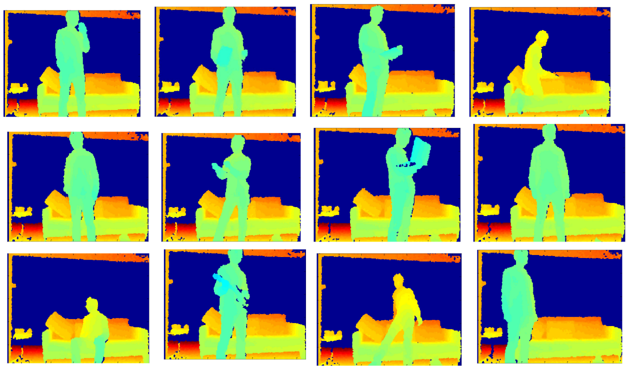
MSR Action Pair3D Dataset (single view)



Method	Input data	Accuracy
Depth Motion Maps [68]	Depth	66.1
Actionlet [58]	Skeleton+Depth	82.2
HON4D [34]	Depth	96.7
SNV [67]	Depth	98.9
Holistic HOPC [39]	Depth	98.3
Local HOPC [40]	Depth	91.7
Ours (HPM)	Depth	66.1
Ours (HPM+TM)	Depth	**99.4**

Figure-Table 2: Sample depth images from the MSR Action Pairs3D dataset [34] containing 6 pairs of actions performed by 10 subjects. Similarities between the poses of action pairs makes this dataset challenging. We used half of the subjects for training and half for testing similar to [34]. We pass the original unsegmented depth frames through our CNN model to extract their view-invariant features. Results are listed in the table above. As expected our HPM achieves low accuracy because each action pair has similar poses and results in similar descriptors through average pooling. Combining our temporal modeling with HPM dramatically improves the accuracy by 33.3%. Our HPM+TM algorithm outperformed all single-view and cross-view methods. It is important to emphasize that single-view based methods [34,58,67,68] exploit the prior knowledge of fixed view point of training and test videos to achieve high accuracy whereas multiview methods do not tune themselves to such prior knowledge or assumption.

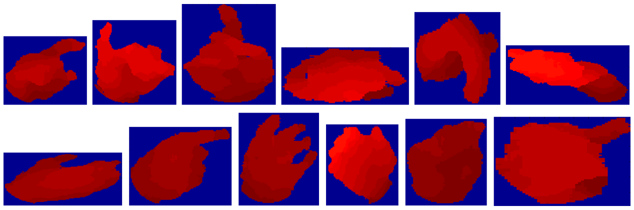
MSR Daily Activity3D dataset (single view)



Method	Input data	Accuracy
AOG [60]	RGB	73.1
BHIM [23]	RGB+Depth	86.9
Interaction Part Mining [74]	RGB+Skeleton	89.3
Actionlet [58]	Skeleton	68.0
LARP [54]	Skeleton	69.4
Actionlet [58]	Skeleton+Depth	85.8
HON4D [34]	Skeleton+Depth	80.0
SNV [67]	Skeleton+Depth	86.3
Holistic HOPC [39]	Skeleton+Depth	88.8
Actionlet [58]	Depth	42.5
Local HOPC [40]	Depth	78.8
Ours (HPM)	Depth	68.1
Ours (HPM+TM)	Depth	80.0
Ours (HPM)	Skeleton+Depth	**95.6**

Figure-Table 3: Sample depth images from the MSR Daily Activity3D dataset [58] containing 16 daily activities performed twice by 10 subjects, once in the standing position and once while sitting. Most activities involve human-object interactions which makes this dataset challenging. Recall that we trained our CNN model using synthetic depth images of only human poses. Following [58], we use videos of half of the subjects for training and half for testing. Results are listed in the table above. HPM alone achieves low accuracy however, combining the proposed temporal modeling with HPM significantly improves the accuracy to 80.0% which is higher than the skeleton only and depth only based methods. For a fair comparison with Skeleton+Depth based methods [34,39,58,67], we combine these features and achieve 95.6% accuracy which is over 6% higher than the nearest competitor.

The MSR Gesture3D dataset (single view)



Method	Input data	Accuracy
Action Graph on Occupancy [26]	Depth	80.5
Action Graph on Silhouette [26]	Depth	87.7
Depth Motion Maps [68]	Depth	89.2
ROP [57]	Depth	88.5
HON4D [34]	Depth	92.5
SNV [67]	Depth	94.7
Holistic HOPC [39]	Depth	**96.2**
Local HOPC [40]	Depth	93.6
Ours (HPM)	Depth	91.0
Ours (HPM+TM)	Depth	94.7

Figure-Table 4: Sample depth images from the MSR Gesture3D dataset [26] which contains 12 American sign language gestures performed 2 to 3 times by 10 subjects. We choose this dataset to show that our learned CNN model is able to generalize to hand gestures even though the CNN model was trained on full human body poses. We use the leave-one-subject-out cross validation scheme [26]. Comparative results are listed in the table above. Actionlet [58], LARP [54] and AOG [60] methods cannot operate on this dataset, because 3D joint positions are not present. Even though our model was trained on fully human body poses, it still competes well with existing methods and achieves the second highest accuracy.

References

Please see the paper for references.