VPN: Learning Video-Pose Embedding for Activities of Daily Living

Supplementary material

Appendix overview
We provide in section 1 computational details regarding the normalization of Euclidean loss provided in Spatial Embedding of RGB and Pose (section 3.2 (II)). Section 2 provides the details of the baseline with LSTM pose backbone with or without coupler in Table 2 & 4 from the ablation studies. Section 3 provides the details of the divergence losses used for comparing with Normalized Euclidean loss in Table 3 from ablation studies. Finally, we provide some more insights about VPN in section 4 to illustrate its effectiveness.
For convenience, we use the same notation as in the main paper for this supplementary material.

1 Details on normalization of Euclidean loss

In equation (4), $\widehat{T_v f_s} = \frac{T_v f_s}{||T_v f_s||_2} = \frac{f_e}{||f_e||_2}$ and $\widehat{T_p z_1} = \frac{T_p z_1}{||T_p z_1||_2} = \frac{P_e}{||P_e||_2}$ are the feature representations projected to the unit hypersphere. Here, we compute the norm $||f_e||_2$ and $||P_e||_2$ using

$$||f_e||_2 = \sqrt{\Sigma_i f_{e_i}^2 + \epsilon} \qquad \& \qquad ||P_e||_2 = \sqrt{\Sigma_i P_{e_i}^2 + \epsilon} \tag{1}$$

where ϵ is a small positive value to prevent dividing zero.

2 LSTM Pose backbone with or without coupler baselines

For the LSTM Pose Backbone in Table 2 & 4, we use a 3-layer stacked LSTM, pre-trained for action classification, as a Pose Backbone by freezing the weights of their cell gates following [2]. The output feature vector h^* is computed by concatenating all the LSTM output features over time. To have a fair comparison with our GCN Pose Backbone, we also introduced residual connections between the original pose input and the LSTM output tensor. However, these residual connections do not improve the action classification accuracy.

For the experiments in Table 2 to implement the attention network without the coupler, we do not perform the step $A_{ST} = inflate(A_S) \circ inflate(A_T)$. Instead, we multiply the attention weights $inflate(A_S)$ and $inflate(A_T)$ separately with the RGB feature map f in two streams following [2]. Finally, the modulated feature maps from both the streams are concatenated to classify the actions.

3 Baselines with KL divergence loss

In Table 3, we compare different forms of KL divergence loss with normalized euclidean loss for spatial embedding of RGB and 3D poses. The KL-divergence losses $D_{KL}(f_e||P_e)$ and $D_{KL}(P_e||f_e)$ for n samples are computed by

$$D_{KL}(f_e||P_e) = \sum_{i=1}^{n} f_e^i log(\frac{f_e^i}{P_e^i}) \tag{2}$$

$$D_{KL}(P_e||f_e) = \sum_{i=1}^{n} P_e^i log(\frac{P_e^i}{f_e^i}) \tag{3}$$

where f_e^i and P_e^i are visual and pose embedding of the i^{th} input sample.
Finally, the bi-directional KL-divergence loss is given by $D_{KL}(f_e||P_e)+D_{KL}(P_e||f_e)$.

4 Detailed qualitative analysis of VPN

In this section, we provide illustrations to show the impact of each VPN components in section 4.1, superiority of VPN compared to other representative baselines in section 4.2, and some result visualization to highlight the solved and remaining challenges in ADL.

4.1 Illustration to show the impact of VPN components

In fig. 1, we illustrate a set of graphs showing the top-5 improvement of action classification accuracy using different components of VPN compared to I3D baseline. As discussed in the ablation studies of the primary paper, each component in VPN is critical for good performance on ADL recognition.

- The spatial embedding provides an accurate alignment of the RGB images and the 3D poses. As a result, the recognition performance of the fine-grained actions improves compared to its counterpart without embedding (see fig. 1 (a)).
- The GCN pose backbone of the attention network, not only provides a strategy to globally optimize the recognition model but also takes the human joint configuration into account for computing the attention weights. This further boosts the action classification performance (see fig. 1 (b)).
- The spatio-temporal coupler of the attention network provides discriminative spatio-temporal attention weights which enables the recognition model to better disambiguate the actions with similar appearance (see fig. 1 (c)).

4.2 Illustration to show the superiority of VPN

We illustrate in fig. 2, the top-5 per-class classification improvement compared to baseline I3D [1] and to an attention mechanism (Separable STA [2]) from the

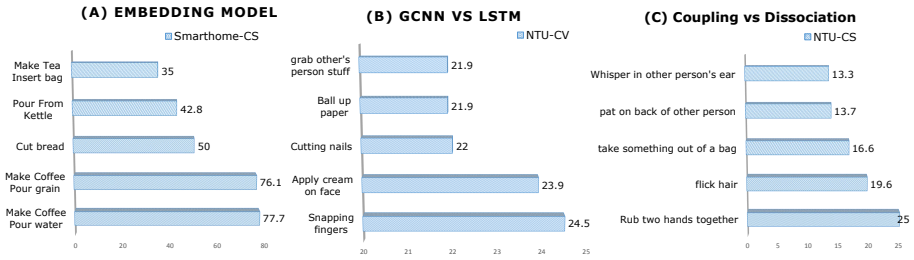

Fig. 1. Graphs illustrating the superiority of each component of VPN compared to their counterparts (without the respective components). We present the Top-5 per class improvement for (a) VPN with embedding vs without embedding (only Spaital Attention), (b) VPN with GCN vs LSTM Pose Backbone, and (c) attention in VPN with vs without spatio-temporal coupler.

state-of-the-art, utilizing 3D poses. The significant accuracy improvements for actions with subtle motion like *hush* (+52.7%), *staple book* (+40.7%) and *reading* (+36.2%) as depicted in fig. 2 (a) illustrate the efficacy of VPN for fine-grained actions. It is worth noting that VPN improves further the classification of actions possessing similar appearance as compared to separable STA in fig. 2 (b). For example, actions like *clapping* (+44.3%) and *flicking hair* (+19.1%) are now discriminated with better accuracy. Further, in fig. 2 (c) we present a radar for the average mis-classification score of few action-pairs. The smaller area under the curve for VPN compared to I3D baseline and Separable STA shows that it is able to better disambiguate the action-pairs even with low inter-class variation.

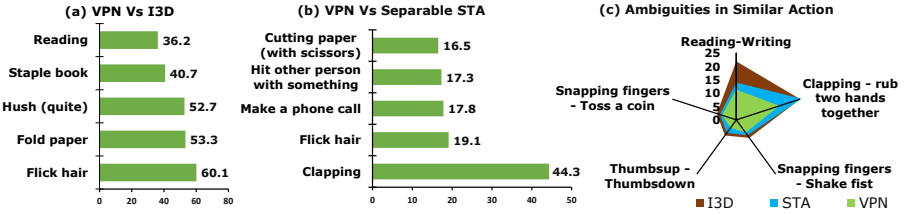

Fig. 2. Graphs illustrating the superiority of VPN compared to the state-of-the-art methods. We present the Top-5 per class improvement for VPN over (a) I3D baseline and (b) Separable STA. In (c), we present a radar for the average mis-classification score of few action-pairs: lower scores indicate lesser ambiguities between the action-pairs.

4.3 Result visualization

In this section, we provide the confusion matrix for action classification on NTU RGB+D 120 and Toyota Smarthome using VPN. In fig 3, we present the confusion matrix of VPN on NTU RGB+D (on right) and a zoom of it around the red

bounding box (on left). We also present the corresponding zoom of the confusion matrix of I3D. We are particularly interested in the mis-classifications performed by VPN and thus, we zoom into the region with relatively low classification accuracy. We observe that actions like *staple book* and *taking something out of bag* were confused with *cutting papers* and *put something into a bag* respectively when classified with I3D. However, with VPN these actions with similar motion are now better discriminated, improving their classification accuracy by approximately 42% and 27% respectively.

Similarly, in fig. 4 (a), we present the confusion matrix of VPN on Toyota Smarthome dataset. In fig. 4 (b), we show the poses for some images belonging to action videos mis-classified by I3D. Thanks to the high quality 3D poses for these videos, now VPN can correctly classify these actions taking the human topology of the 3D poses into account. We provide some visual results in fig. 5 where VPN outperforms I3D baseline. We notice that actions like *Drink from glass* are not recognized due to extremely low number of training samples. We further notice that actions like *using tablet* are recognized with low accuracy of 13% and largely confused with *using laptop*. However, I3D completely misclassifies the action *using tablet*. We also observe that still few action classes are recognized with extremely low classification accuracy. We infer that these poor classification results on certain videos are due to occlusion, low resolution of the actions and low quality poses as illustrated in fig 6.

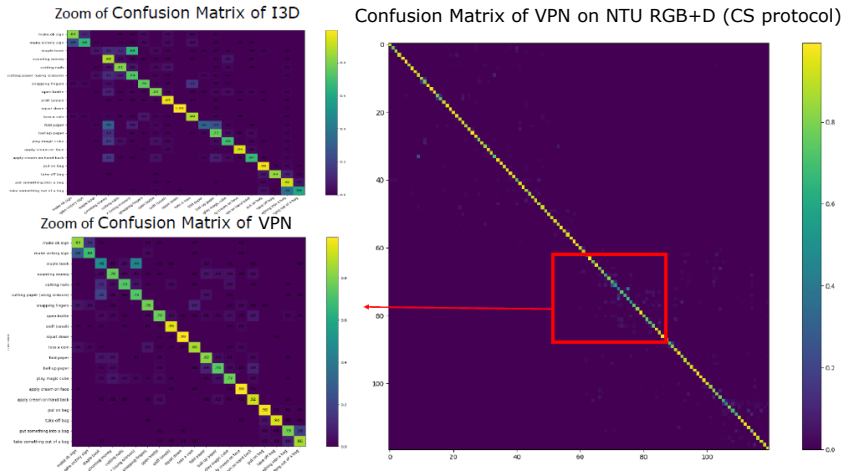

Fig. 3. Confusion matrix of VPN on NTU RGB+D (CS Protocol) on the right. Zoom of the red bounding box on the left along with the corresponding confusion matrix of I3D.

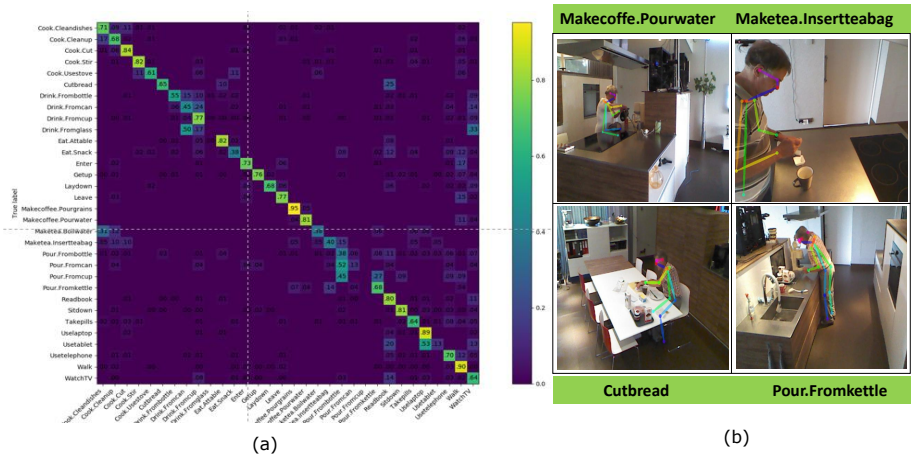

Fig. 4. (a) Confusion matrix of VPN on Toyota Smarthome (CS protocol) (b) Illustration of poses for activities mis-classified with I3D but correctly classified with VPN.

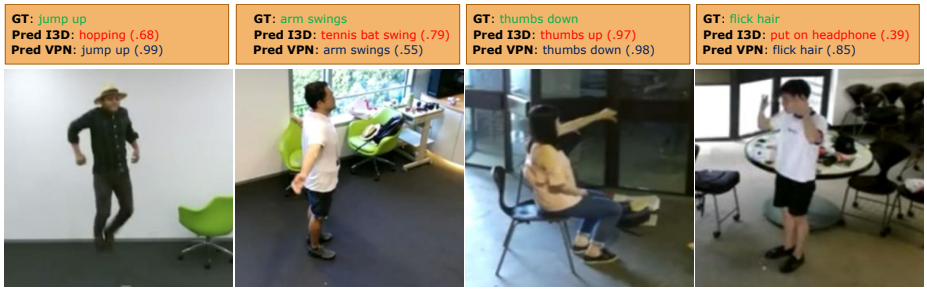

Fig. 5. Visual results from NTU RGB+D 120 where VPN outperforms I3D.

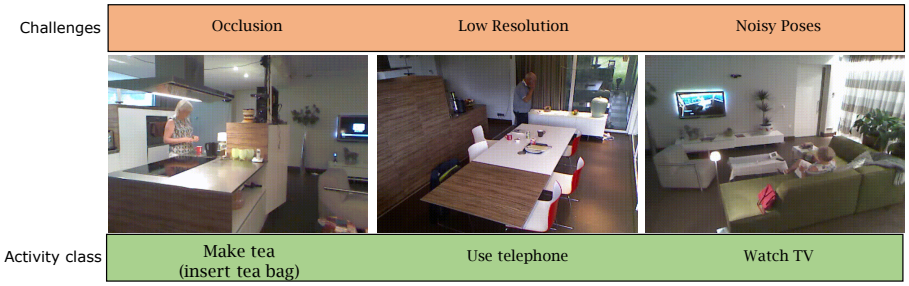

Fig. 6. Illustration of the remaining challenges in Toyota Smarthome with images from activities (indicated below) and their corresponding challenges (indicated on the top)

References

1. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4724–4733. IEEE (2017)
2. Das, S., Dai, R., Koperski, M., Minciullo, L., Garattoni, L., Bremond, F., Francesca, G.: Toyota smarthome: Real-world activities of daily living. In: ICCV (2019)