Supplemental: Multi-Label Activity Recognition using Activity-specific Features and Activity Correlations

Yanyi Zhang,[1] Xinyu Li, [1,2] Ivan Marsic [1]

[1] Rutgers University–New Brunswick, Electrical and Computer Engineering Department
[2] Amazon Web Services

1. Experimental Results on Hockey

We also evaluate our method on the Hockey dataset besides the Charades and the Volleyball datasets in the main paper. The Hockey dataset [2] was collected from real university-level hockey matches using two fixed cameras positioned at both ends of the rink on the spectator's side. It includes 12 multi-label activities in 36 videos (each video contains around 10000 frames).

Based on the accuracy and F1. scores on Table 1, our system substantially outperformed all the existing approaches on Hockey for multi-label activity recognition. We also compared our method with the baseline using the latest activity recognition model (CSN-152 baseline in Table 1) due to the other existing approaches on Hockey are relatively old [4]. Our model roughly higher 2.5% F1. score compared to the baseline network, which shows that the activity-specific features also improve the performance on the Hockey dataset.

Table 1. Experimental results for multi-label activity recognition on Hockey.

Hockey		
method	Acc.	F1.
SVR [2]	-	16.0
EO-SVM [1]	90.0	-
CNN Over time [3]	-	42.0
CSN-152 baseline [4]	95.2	57.1
Our's	**96.3**	**60.2**

2. Backbone

We use CSN-152 [4] as the backbone of our model. We remove the temporal strides in res_5 to make the backbone features F_f contain more complete information from the original video. Table 2 shows the detail parameters of the backbone network and the parameters for generating the $Attn_O$ and $Attn_A$ in our proposed network.

Table 2. The detail structure and parameters of the i3D network that we are using.

Stage	Details	Output Size
Conv_1	$5 \times 7 \times 7$, 64, stride 1, 2, 2	$32 \times 112 \times 112 \times 64$
Maxpool_1	$2 \times 3 \times 3$, stride 2, 2, 2	$16 \times 56 \times 56 \times 64$
Res_2	$\begin{pmatrix} 3 \times 1 \times 1, 64 \\ 1 \times 3 \times 3, 64 \\ 1 \times 1 \times 1, 256 \end{pmatrix} \times 3$	$16 \times 56 \times 56 \times 256$
Maxpool_2	$2 \times 3 \times 3$, stride 2, 2, 2	$8 \times 28 \times 28 \times 256$
Res_3	$\begin{pmatrix} 3 \times 1 \times 1, 128 \\ 1 \times 3 \times 3, 128 \\ 1 \times 1 \times 1, 512 \end{pmatrix} \times 4$	$8 \times 28 \times 28 \times 512$
Res_4	$\begin{pmatrix} 3 \times 1 \times 1, 256 \\ 1 \times 3 \times 3, 256 \\ 1 \times 1 \times 1, 1024 \end{pmatrix} \times 23$	$8 \times 14 \times 14 \times 1024$
Res_5	$\begin{pmatrix} 3 \times 1 \times 1, 512 \\ 1 \times 3 \times 3, 512 \\ 1 \times 1 \times 1, 2048 \end{pmatrix} \times 3$	$8 \times 7 \times 7 \times 2048$
$Attn_O$	$64, 392 \times 392, 128$	64×128
$Attn_A$	$64 \times A, 128$	$A \times 128$
FC	128×1	$1 \times A$

3. Visualizing Activity-specific features

We further visualized activity-specific feature maps for more videos in Charades. Figure 1, Figure 2 and Figure 3 are the visualized activity-specific features in 7 different video clips. The activity-specific features maps and the bounding boxes on their corresponding input frames are generated using the same method proposed in Section 6 (Feature Visualization) of the main paper.

References

[1] Marc-André Carbonneau. *Multiple instance learning under real-world conditions*. PhD thesis, École de technologie supérieure, 2017.

[2] Marc-André Carbonneau, Alexandre J Raymond, Eric Granger, and Ghyslain Gagnon. Real-time visual play-break detection in sport events using a context descriptor. In *2015 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 2808–2811. IEEE, 2015.

[3] Konstantin Sozykin, Stanislav Protasov, Adil Khan, Rasheed Hussain, and Jooyoung Lee. Multi-label class-imbalanced action recognition in hockey videos via 3d convolutional neural networks. In *2018 19th IEEE/ACIS International Conference*

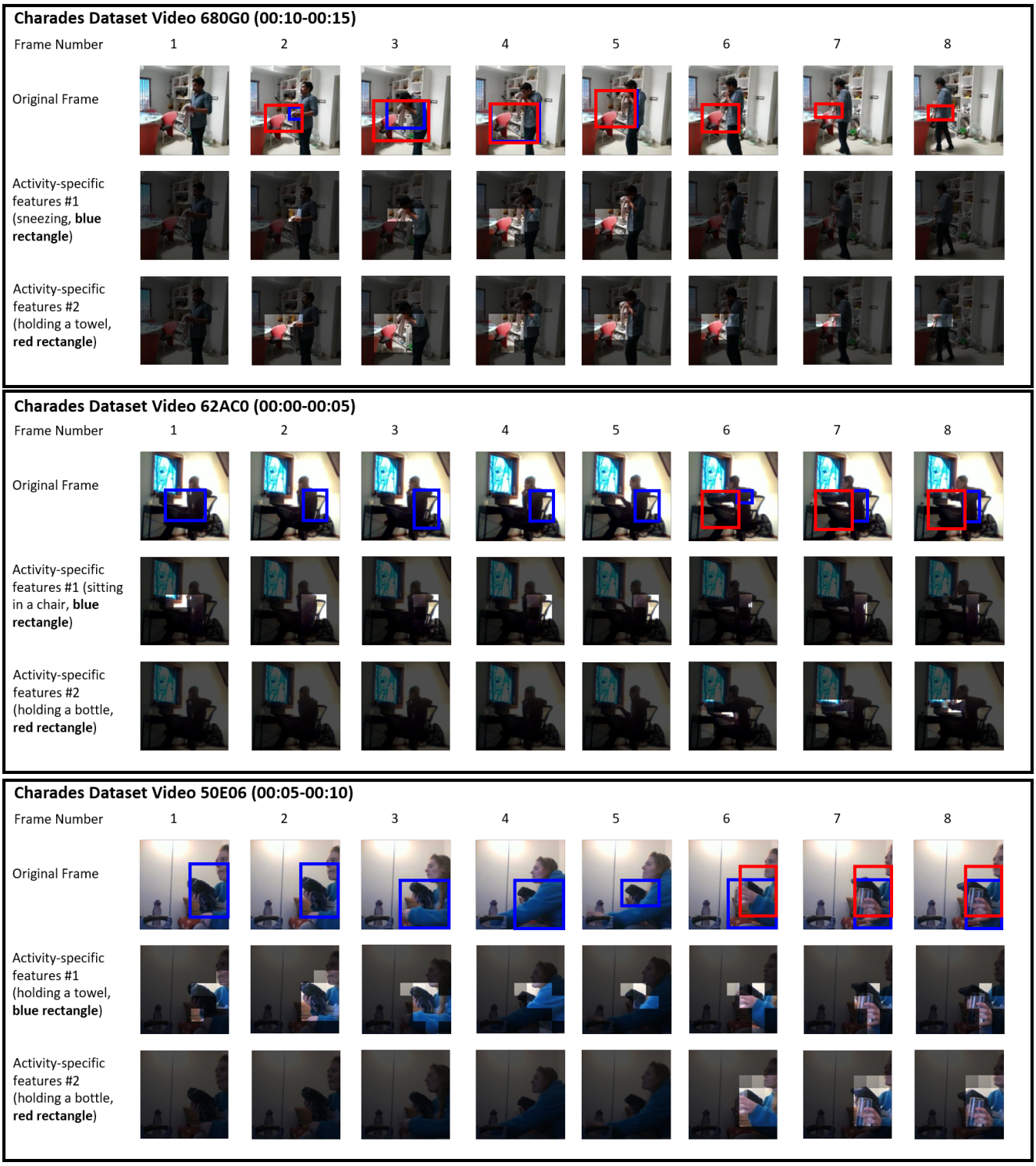

Figure 1. Visualization of activity-specific features in video "680G0", "62ACD" and "50E06" from Charades dataset. Detail explanation is in Section 6 of the main paper

on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), pages 146–151. IEEE, 2018.

[4] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *Proceedings of the IEEE International Conference*

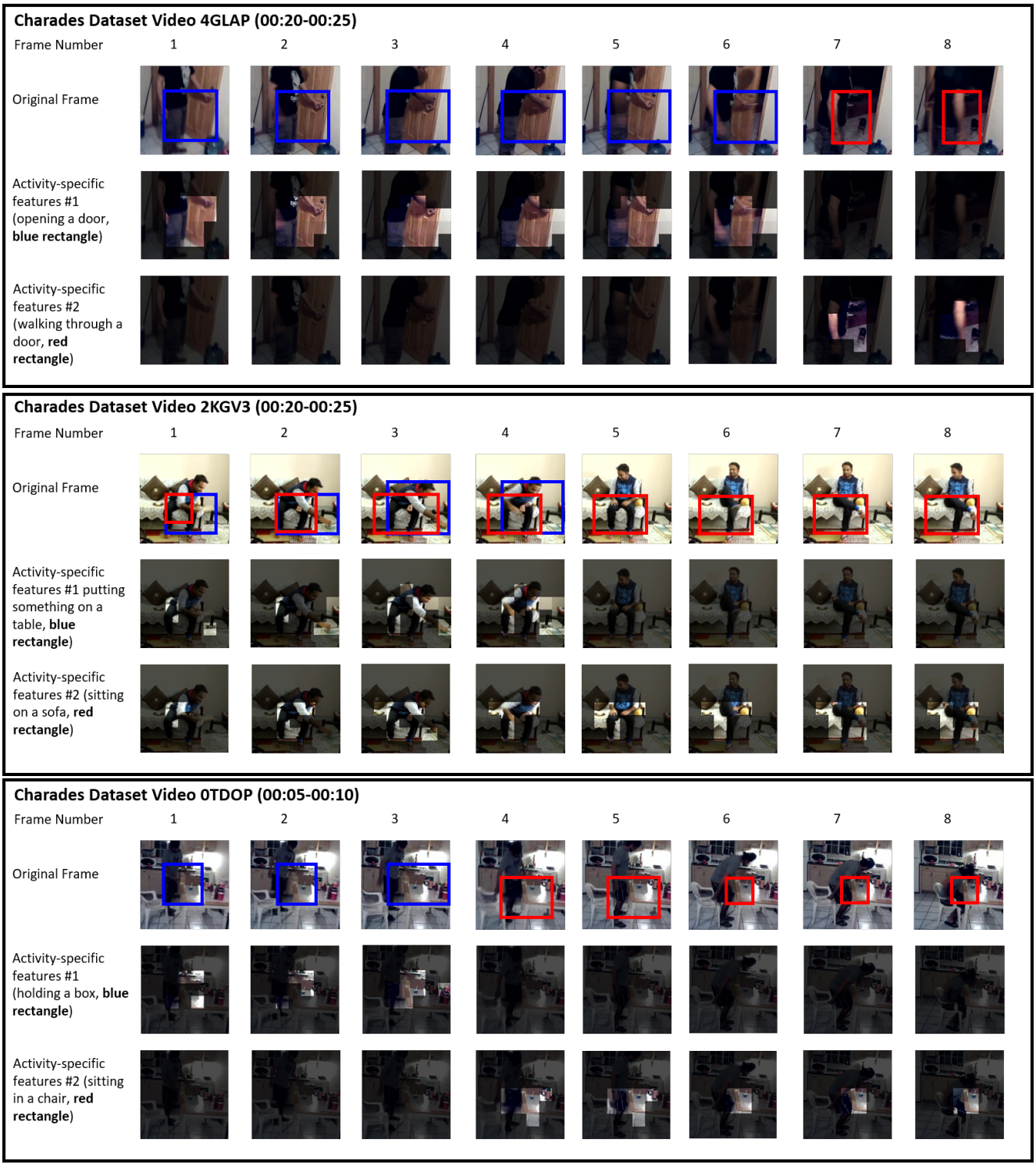

Figure 2. Visualization of activity-specific features in video "4GLAP", "2KGV3" and "0TDOP" from Charades dataset. Detail explanation in Section 5.1 of the main paper

on Computer Vision, pages 5552–5561, 2019.

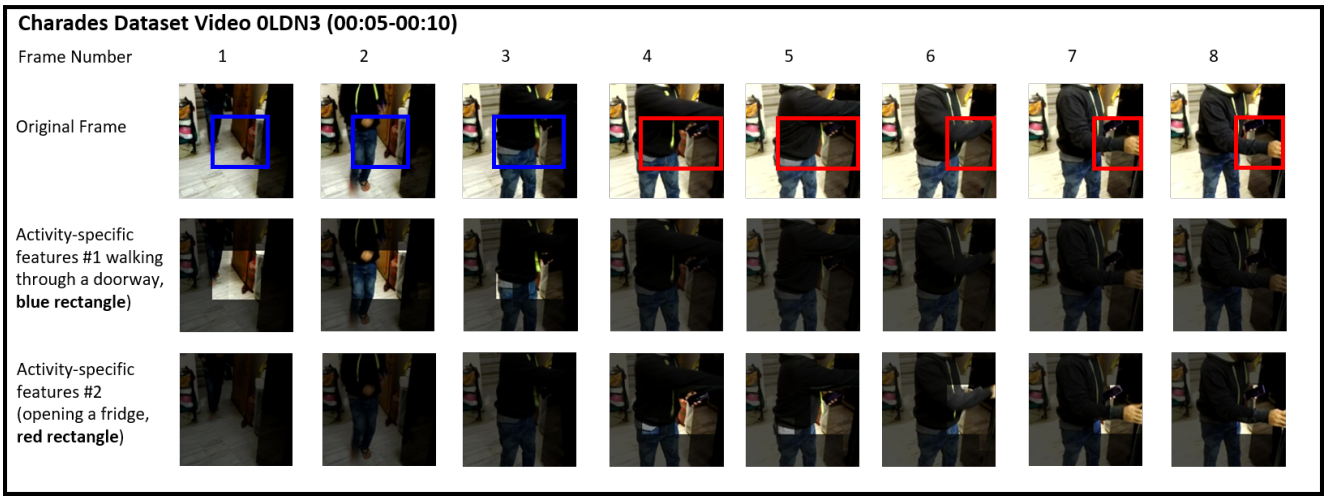

Figure 3. Visualization of activity-specific features in video "0LDN3" from Charades dataset. Detail explanation in Section 5.1 of the main paper