Supplementary Material: Compressed Video Action Recognition

Chao-Yuan Wu[1,5] cywu@cs.utexas.edu

Manzil Zaheer[2,5] manzil@cmu.edu

Hexiang Hu[3,5] hexiangh@usc.edu

R. Manmatha[4] manmatha@a9.com

Alexander J. Smola[5] smola@amazon.com

Philipp Krähenbühl[1] philkr@cs.utexas.edu

[1]The University of Texas at Austin, [2]Carnegie Mellon University,
[3]University of Southern California, [4]A9, [5]Amazon

1. RNN-Based Models

Given the recurrent definition of P-frames, one can use a RNN to model a compressed video. In preliminary experiments, we experiment with a variant using Conv-LSTMs [4].

The architecture is identical to CoViAR except that i) it uses the original $\mathcal{T}$ and Δ instead of the accumulated $\mathcal{D}$ and $\mathcal{R}$, because here we want to the original dependency, and ii) it uses a Conv-LSTM to aggregate the CNN features instead of average pooling. Formally, let $x^{(t)}_{\text{fusion}} := \max\left(x^{(t)}_{\text{motion}}, x^{(t)}_{\text{residual}}\right)$ denote the max-pooled P-frame feature at time t. The Conv-LSTM takes the input sequence

$$\left(x^{(0)}_{\text{RGB}}, x^{(1)}_{\text{fusion}}, x^{(2)}_{\text{fusion}}, \dots\right).$$

Here the number of channels of $x^{(0)}_{\text{RGB}}$ is reduced from 2048 to 512 by an 1×1 convolution so that its dimensionality matches $x^{(t)}_{\text{fusion}}$. We use 512-dimensional hidden states and 3×3 kernels for the Conv-LSTM. Due to memory constraint, we subsample one every two P-frames to reduce the sequence length.

Table 1 presents the results. Even though the Conv-LSTM model outperforms traditional RGB-based methods, the decoupled CoViAR achieves the best performance. We also try adding the input of Conv-LSTM to its output as a skip connection, but it leads to worse performance (Conv-LSTM-Skip).

RGB-only	Conv-LSTM	Conv-LSTM-Skip	CoViAR
88.4	<u>89.1</u>	87.8	**90.8**

Table 1: Accuracy on UCF-101 split 1. CoViAR decouples the long dependency and outperforms RNN-based models.

2. Feature Fusion

We experiment with different ways of combining P-frame features, $x^{(t)}_{\text{motion}}$, $x^{(t)}_{\text{residual}}$, and I-frame features $x^{(0)}_{\text{RGB}}$. In particular, we evaluate maximum, mean, and multiplicative fusion, concatenation of feature maps, and late fusion (summing softmax scores). For maximum, mean, and multiplicative fusion, we perform 1×1 convolution on I-frame feature maps before fusion, so that their dimensionality matches P-frame features.

Table 2 summarizes the results; we found late fusion works the best for CoViAR. Note that late fusion allows training of a decoupled model, while the rest requires training multiple CNNs jointly. The ease of training of late fusion may also contribute to its superior performance.

Max	Mean	Mult	Concat	Late
87.9	88.1	87.8	<u>89.7</u>	**90.8**

Table 2: Accuracy on UCF-101 split 1 with different feature fusion methods.

3. CoViAR without Temporal Segments

For further analysis, we also evaluate CoViAR without using temporal segments [3] (Table 3). It still significantly outperforms models using RGB images only, including ResNet-152 (83.4% in ST-Mult [1]; 84.7% with out implementation) and Res3D [2] (85.8%).

I	M	R	I+M	I+R	I+M+R
84.7	63.4	76.6	87.9	87.2	**88.9**

Table 3: Accuracy of CoViAR without temporal segments on UFC-101 split 1.

4. Confusion Matrix

Figure 1 and Figure 2 show the confusion matrices of CoViAR and the model using only RGB images respectively, on UCF-101. Figure 3 shows the difference between their predictions. We can see that CoViAR corrects many mistakes made by the RGB-based model (off-diagonal purple blocks in Figure 3). For example, while the RGB-based model gets confused about the similar actions of *Cricket Bowling* and *Cricket Shot*, our model better distinguishes between them.

References

[1] C. Feichtenhofer, A. Pinz, and R. P. Wildes. Spatiotemporal multiplier networks for video action recognition. In *CVPR*, 2017. 1

[2] D. Tran, J. Ray, Z. Shou, S.-F. Chang, and M. Paluri. ConvNet architecture search for spatiotemporal feature learning. *arXiv preprint arXiv:1708.05038*, 2017. 1

[3] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016. 1

[4] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *NIPS*, 2015. 1

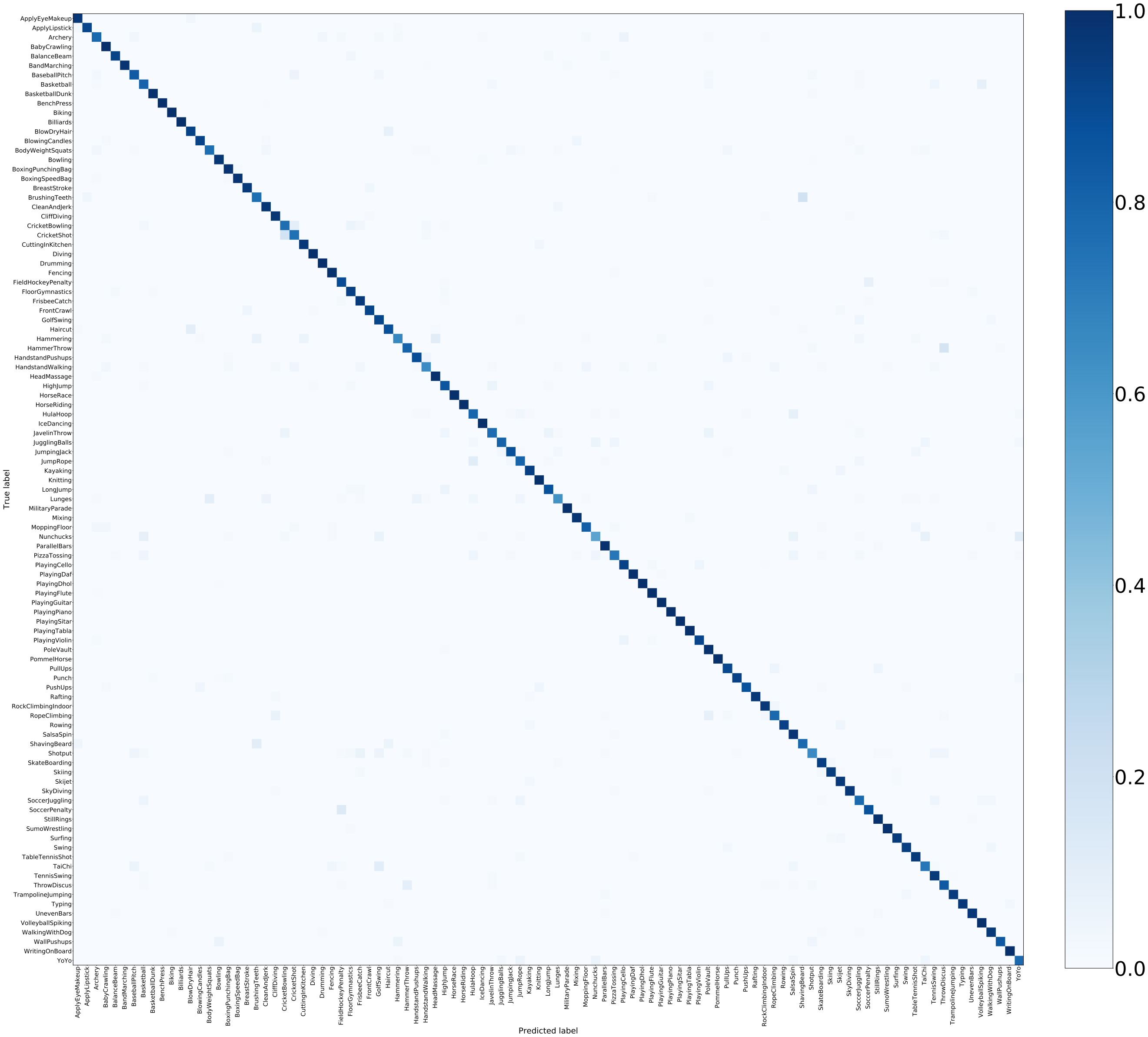

Figure 1: Confusion matrix of CoViAR on UCF-101.

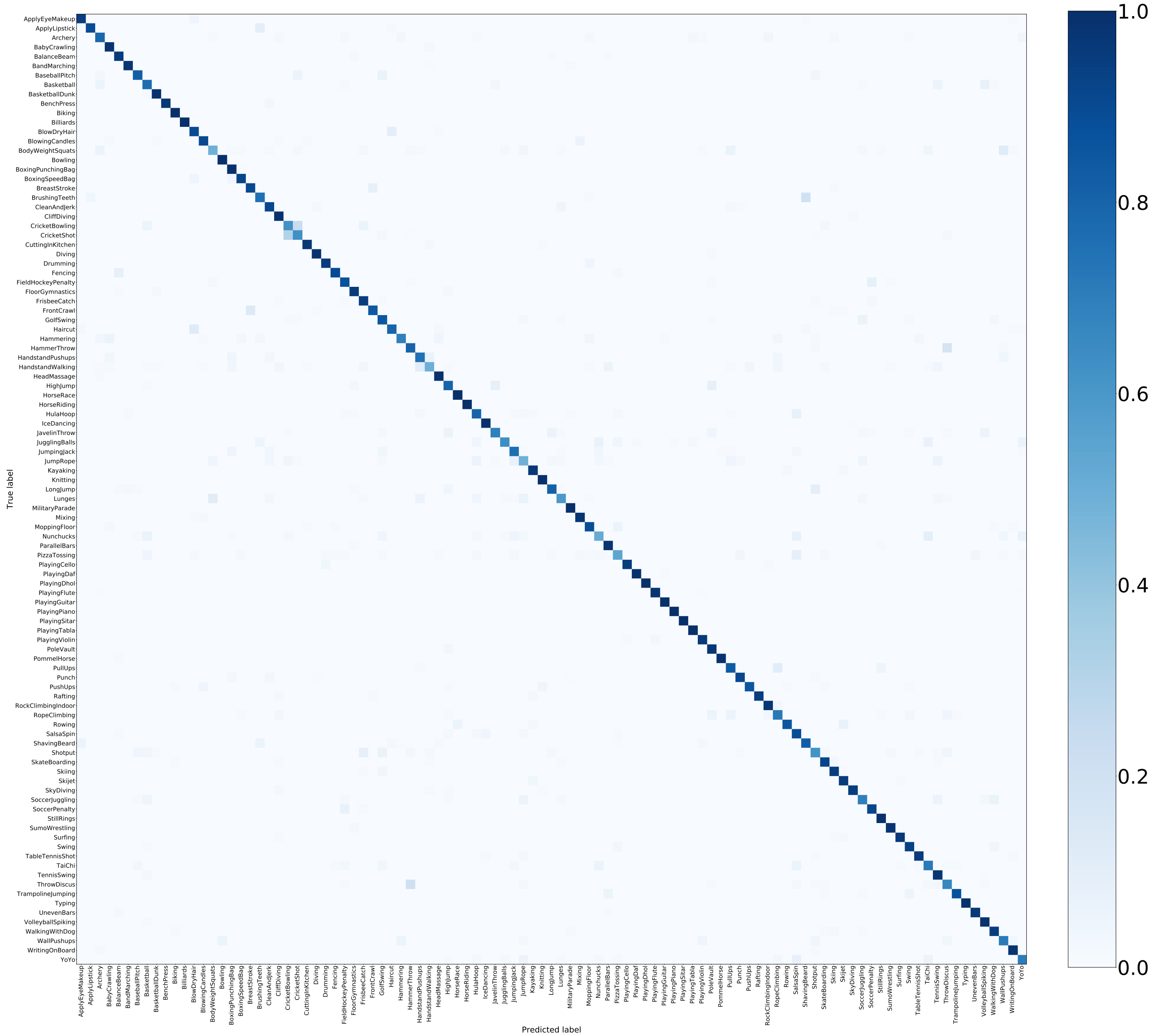

Figure 2: Confusion matrix of the model using RGB images on UCF-101.

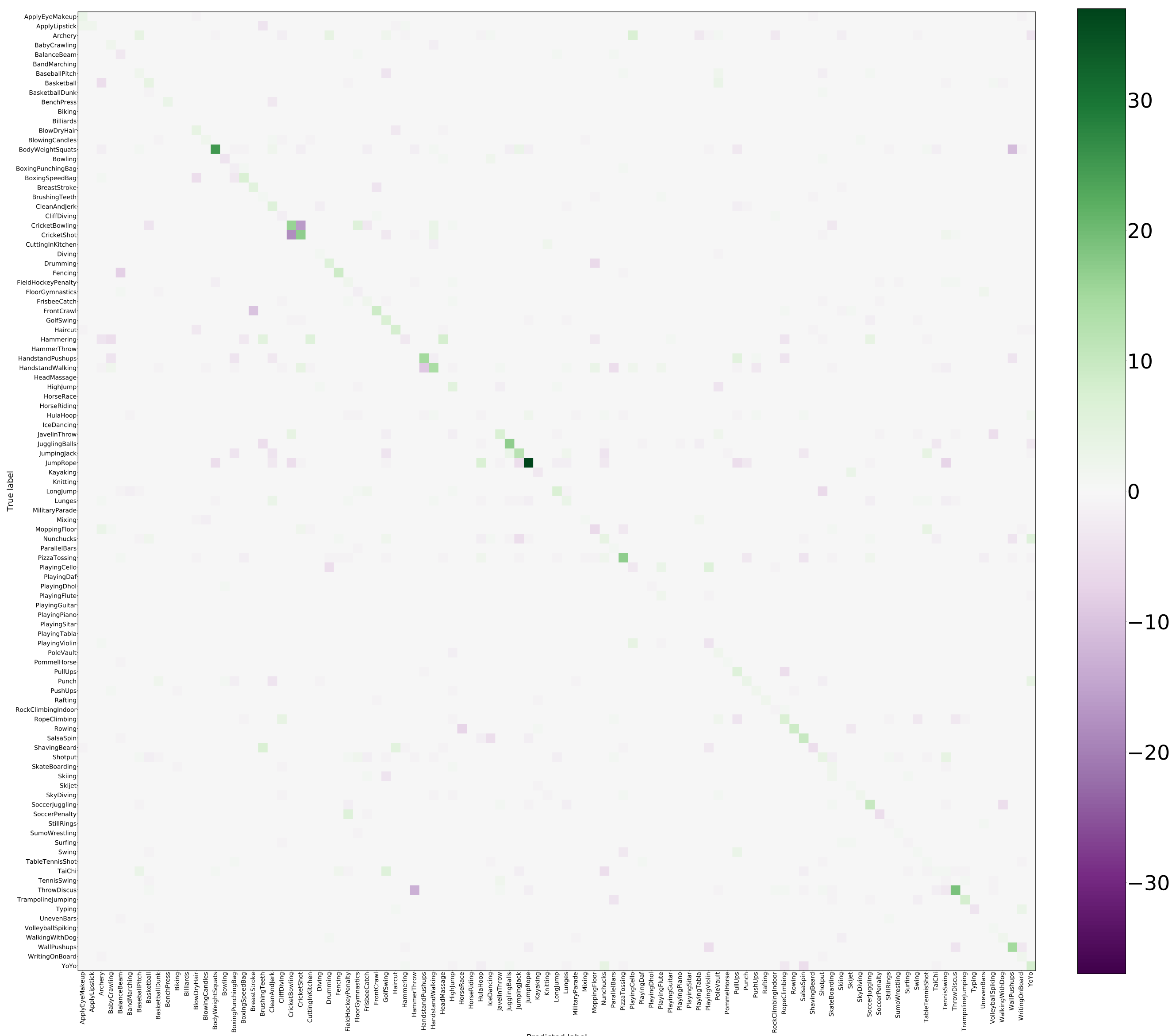

Figure 3: Difference between CoViAR's predictions and the RGB-based model's predictions. For diagonal entries, positive values (in green) is better (increase of correct predictions). For off-diagonal entries, negative values (purple) is better (reduction of wrong predictions).