Human Action Segmentation using 3D Fully Convolutional Network Supplementary Material

Pei Yu[1]
pyi980@eecs.northwestern.edu
Jiang Wang[2]
wangjiang@google.com
Ying Wu[1]
yingwu@northwestern.edu

[1] Northwestern University
2145 Sheridan Road
Evanston, IL, USA

[2] Google Research
1600 Amphitheatre Pkwy,
Mountain View, CA, USA

1 Derivation of Equations in Sec. 3.5

Denote the spatio-temporal observation at pixel i and time t as $z_{t,i}$. Denote the action label at pixel i and time t as $l_{t,i}$. The dynamics of $l_{t,i}, t = \{1,\dots,T\}$ can be modeled as a typical HMM model. State transition probability $P(l_{t+1,i}|l_{t,i})$ can be estimated from ground truth labeling. We use the same state transition probability for all pixel locations.

The output softmax probability of 3DFCN provides posterior $P(l_{t,i}|z_{t,i})$ based on spatio-temporal observation $z_{t,i}$ at time t. With T frames of a video, belief of hidden state based on all observations is $P(l_{t,i}|\underline{Z}_{T,i})$, where $\underline{Z}_{t,i} = \{z_{1,i},\dots,z_{t,i}\}$, $\bar{Z}_{t,i} = \{z_{t+1,i},\dots,z_{T,i}\}$.

Derivation of equation (1) in main submission.

$$
\begin{aligned}
P(l_{t,i}|\underline{Z}_{T,i}) &= \frac{P(\underline{Z}_{t,i},\bar{Z}_{t,i}|l_{t,i})P(l_{t,i})}{P(\underline{Z}_{T,i})} \\
&= \frac{P(l_{t,i})P(\underline{Z}_{t,i}|l_{t,i})P(\bar{Z}_{t,i}|l_{t,i})}{P(\underline{Z}_{T,i})} \\
&= \frac{P(\underline{Z}_{t,i})P(\bar{Z}_{t,i})P(l_{t,i}|\underline{Z}_{t,i})P(l_{t,i}|\bar{Z}_{t,i})}{P(\underline{Z}_{T,i})P(l_{t,i})} \\
&\propto \frac{P(l_{t,i}|\underline{Z}_{t,i})P(l_{t,i}|\bar{Z}_{t,i})}{P(l_{t,i})}
\end{aligned}
\tag{1}
$$

Derivation of equation (2) in main submission.

$$
\begin{aligned}
P(l_{t,i}|\underline{Z}_{t,i}) &= \frac{P(l_{t,i},\underline{Z}_{t-1,i},z_{t,i})}{P(\underline{Z}_{t,i})} \\
&= \frac{\int_{l_{t-1,i}} P(l_{t,i},z_{t,i}|\underline{Z}_{t-1,i},l_{t-1,i})P(\underline{Z}_{t-1,i},l_{t-1,i})}{P(\underline{Z}_{t,i})} \\
&= \frac{\int_{l_{t-1,i}} P(l_{t,i},z_{t,i}|l_{t-1,i})P(\underline{Z}_{t-1,i},l_{t-1,i})}{P(\underline{Z}_{t,i})} \\
&= \frac{P(z_{t,i}|l_{t,i})\int_{l_{t-1,i}} P(l_{t,i}|l_{t-1,i})P(\underline{Z}_{t-1,i},l_{t-1,i})}{P(\underline{Z}_{t,i})} \\
&= \frac{P(z_{t,i}|l_{t,i})}{P(\underline{Z}_{t,i})}\int_{l_{t-1,i}} P(l_{t,i}|l_{t-1,i})P(\underline{Z}_{t-1,i})P(l_{t-1,i}|\underline{Z}_{t-1,i}) \\
&= \frac{P(\underline{Z}_{t-1,i})P(z_{t,i}|l_{t,i})}{P(\underline{Z}_{t,i})}\int_{l_{t-1,i}} P(l_{t,i}|l_{t-1,i})P(l_{t-1,i}|\underline{Z}_{t-1,i}) \\
&= \frac{P(\underline{Z}_{t-1,i})P(l_{t,i}|z_{t,i})P(z_{t,i})}{P(\underline{Z}_{t,i})P(l_{t,i})}\int_{l_{t-1,i}} P(l_{t,i}|l_{t-1,i})P(l_{t-1,i}|\underline{Z}_{t-1,i}) \\
&\propto \frac{P(l_{t,i}|z_{t,i})}{P(l_{t,i})}\int_{l_{t-1,i}} P(l_{t,i}|l_{t-1,i})P(l_{t-1,i}|\underline{Z}_{t-1,i})
\end{aligned}
\tag{2}
$$

Derivation of equation (3) in main submission.

$$
\begin{aligned}
&P(l_{t,i}|\bar{Z}_{t,i}) \\
&= \int_{l_{t+1,i}} \frac{P(l_{t,i},z_{t+1,i}|l_{t+1,i},\bar{Z}_{t+1,i})P(l_{t+1,i},\bar{Z}_{t+1,i})}{P(\bar{Z}_{t,i})} \\
&= \frac{1}{P(\bar{Z}_{t,i})}\int_{l_{t+1,i}} P(l_{t,i},z_{t+1,i}|l_{t+1,i},\bar{Z}_{t+1,i})P(\bar{Z}_{t+1,i})P(l_{t+1,i}|\bar{Z}_{t+1,i}) \\
&= \frac{P(\bar{Z}_{t+1,i})}{P(\bar{Z}_{t,i})}\int_{l_{t+1,i}} P(l_{t,i},z_{t+1,i}|l_{t+1,i})P(l_{t+1,i}|\bar{Z}_{t+1,i}) \\
&= \frac{P(\bar{Z}_{t+1,i})}{P(\bar{Z}_{t,i})}\int_{l_{t+1,i}} P(z_{t+1,i}|l_{t,i},l_{t+1,i})P(l_{t,i}|l_{t+1,i})P(l_{t+1,i}|\bar{Z}_{t+1,i}) \\
&= \frac{P(\bar{Z}_{t+1,i})}{P(\bar{Z}_{t,i})}\int_{l_{t+1,i}} P(z_{t+1,i}|l_{t+1,i})P(l_{t,i}|l_{t+1,i})P(l_{t+1,i}|\bar{Z}_{t+1,i}) \\
&= \frac{P(\bar{Z}_{t+1,i})}{P(\bar{Z}_{t,i})}\int_{l_{t+1,i}} \frac{P(l_{t+1,i}|z_{t+1,i})P(z_{t+1,i})P(l_{t+1,i}|l_{t,i})P(l_{t,i})}{P^2(l_{t+1,i})}P(l_{t+1,i}|\bar{Z}_{t+1,i}) \\
&= \frac{P(\bar{Z}_{t+1,i})P(z_{t+1,i})P(l_{t,i})}{P(\bar{Z}_{t,i})}\int_{l_{t+1,i}} \frac{P(l_{t+1,i}|z_{t+1,i})P(l_{t+1,i}|l_{t,i})}{P^2(l_{t+1,i})}P(l_{t+1,i}|\bar{Z}_{t+1,i}) \\
&\propto P(l_{t,i})\int_{l_{t+1,i}} \frac{P(l_{t+1,i}|z_{t+1,i})P(l_{t+1,i}|l_{t,i})}{P^2(l_{t+1,i})}P(l_{t+1,i}|\bar{Z}_{t+1,i})
\end{aligned}
\tag{3}
$$

Method	Recall		Precision	
	Foreground	Localization	Foreground	Localization
3DFCN-32s	**92.8**	**47.7**	48.9	25.1
3DFCN-8s	80.3	40.6	62.9	31.8
3DFCN-8s + HMM	84.9	42.4	58.9	29.4
3DFCN-8s + CRFs	70.8	38.2	**72.1**	**38.9**
3DFCN-8s + HMM + CRFs	81.3	42.3	64.5	33.5
2DFCN-8s	71.4	19.4	36.8	10.0
2DFCN-8s + HMM	82.6	21.7	32.0	8.4
2DFCN-8s + CRFs	42.3	14.9	50.2	17.7
2DFCN-8s + HMM + CRFs	71.6	20.4	39.4	11.2
[1]	72.0	-	63.0	-

Table 1: Performance of proposed methods and baseline methods.

2 Experimental Results

2.1 Qualitative results of different sampling stride lengths

In this section, we provide the qualitative results of different network sampling stride lengths. In particular, the results of 3DFCN-32s and 3DFCN-8s are provided in Fig. 1 to demonstrate the impact of reduced sampling stride length, which is achieved via dilation. As shown in Fig. 1, with lower network sampling stride length, 3DFCN-8s achieves better segmentation accuracy than 3DFCN-32s.

2.2 Results of each action class

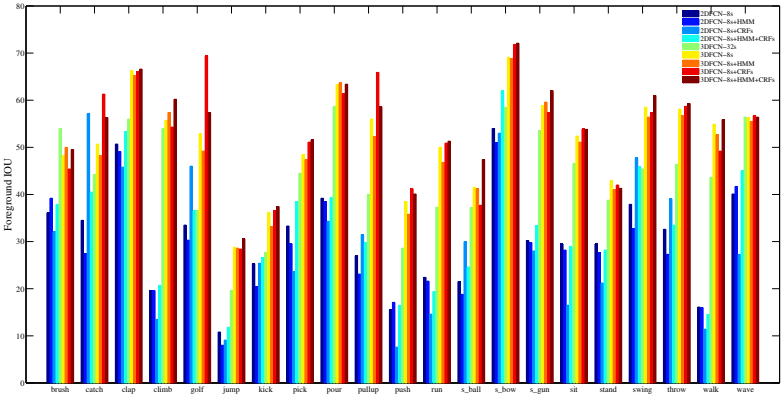
In this section, we provide the experimental results for each action class out of total 21 action classes in J-HMDB. Besides IOU, we also provide the performance of recall and precision. Similar to action foreground IOU and localization IOU, the recall and precision metrics include foreground recall and precision, and localization recall and precision. For foreground recall and precision, the pixels of all action classes are treated as action foreground. For localization recall and precision, the metric is calculated for each action class. Note that in [1], the reported IOU performance on J-HMDB is only action foreground IOU. Action localization IOU, recall and precision are not reported in [1].

The performance of foreground IOU, recall and precision are summarized in Fig. 2. The performance of localization IOU, recall and precision are summarized in Fig. 3. It is shown that compared with 2DFCN-based methods, 3DFCN-based methods generally achieve superior IOU performance on each action class. 2DFCN-based methods achieve foreground recall performance comparable to 3DFCN-based methods. However, the performance of foreground precision is low. For the performance of localization IOU, recall and precision, 3DFCN-based methods outperform 2DFCN-based methods, since 2DFCN-based methods are inferior on inferring action class label. Average foreground recall, foreground precision, localization recall and localization precision are summarized in Tab. 1.

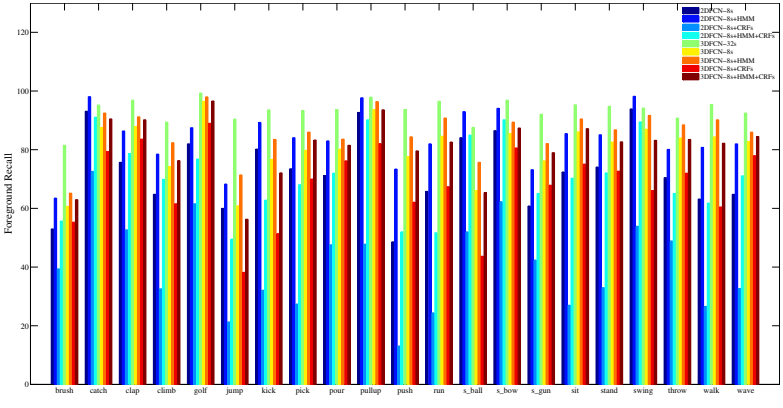
References

[1] Jiasen Lu, Jason J Corso, et al. Human action segmentation with hierarchical supervoxel consistency. In *CVPR*, pages 3762–3771, 2015.

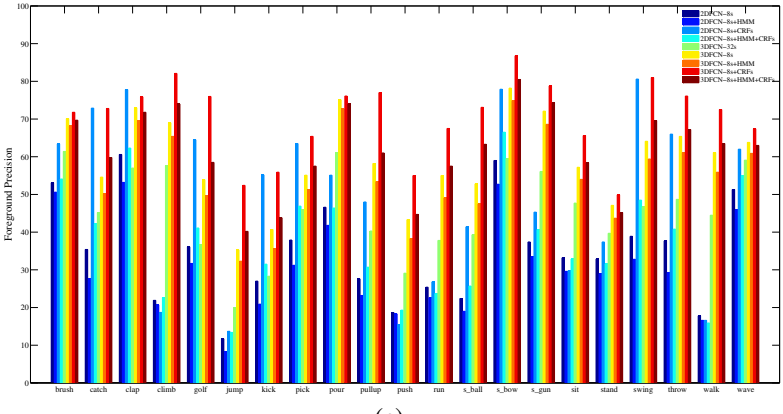
Figure 1: Comparison of different sampling stride lengths. Each column contains the results of one out of total 21 action classes. For each action class, the first row is ground truth action segmentation. The second row is the result of 3DFCN-32s. The third row is the result of 3DFCN-8s.



(a)

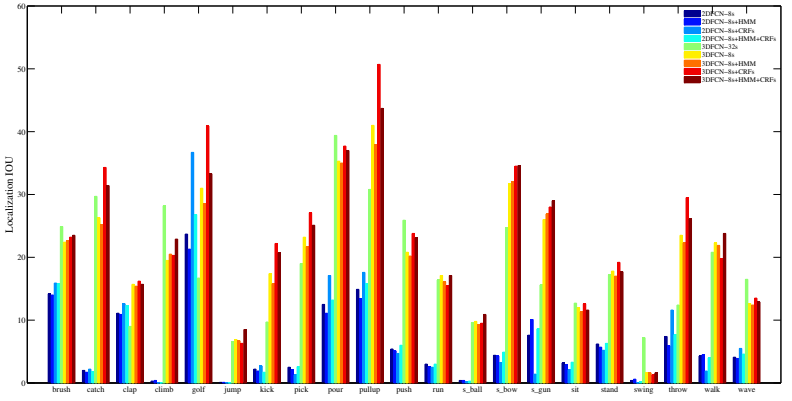


(b)

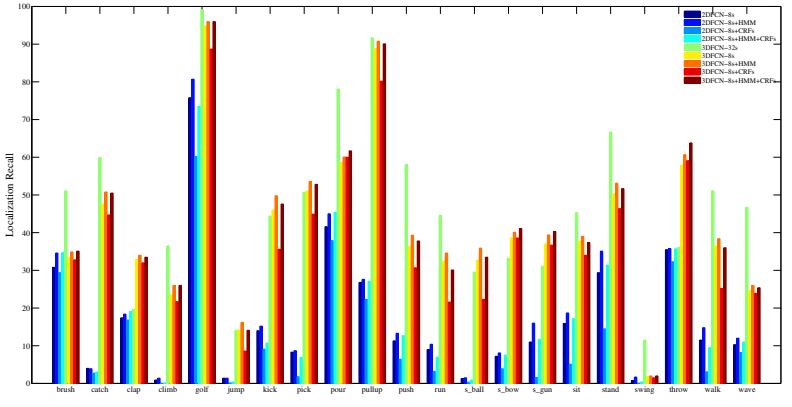


(c)

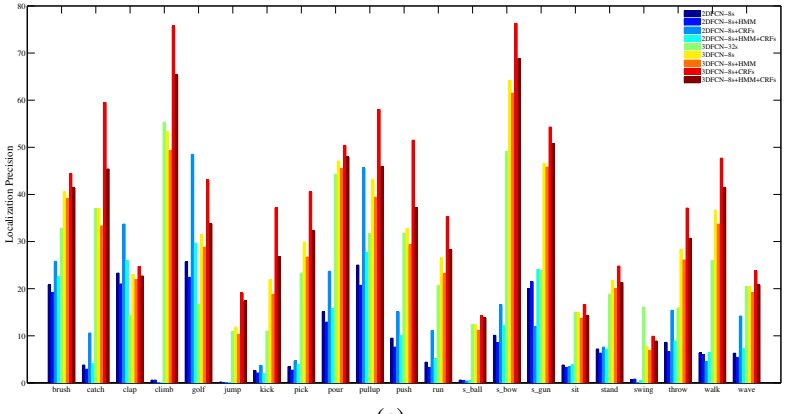
Figure 2: Results of foreground IOU, foreground recall and foreground precision of each action class.



(a)



(b)



(c)

Figure 3: Results of localization IOU, localization recall and localization precision of each action class.