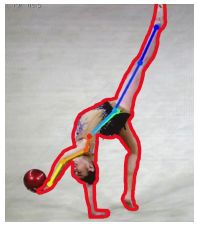
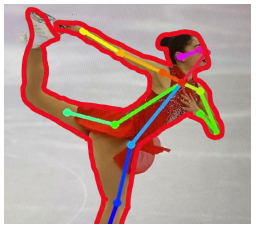
Dilated Silhouette Convolutional Neural Network for Human Action Recognition (#1314)

Problem

- Human action is a spatio-temporal motion sequence with strong inter-dependencies between spatial geometry and temporal dynamics
- Current recognition algorithms lack synergy in investigating space and time in a joint representation and embedding space
- Recognition suffers from view changes, camera motion, background clutter, occlusion, anthropometry, and variation in action execution rate

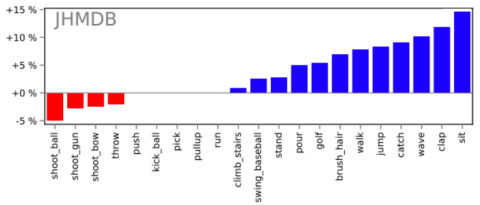




Current geometry-based methods estimate the skeleton and use it for recognition. However, the skeletons are often inaccurate while the silhouettes are robust.

Findings

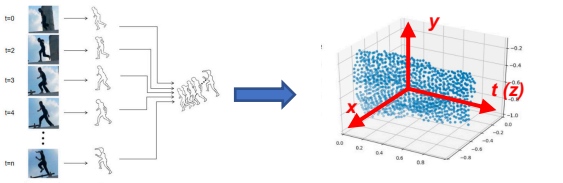
Methods	JHMDB	HMDB	UCF101
2D geometry-based methods			
P-CNN Chéron et al. (2015)	61.1	-	-
Action Tubes Gkioxari and Malik (2015)	62.5	-	-
PoTion Choutas et al. (2018)	57.0	43.7	65.2
PA3D Yan et al. (2019)	69.5	55.3	-
DD-Net Zhang et al. (2019)	77.2	-	-
SCN (Ours)	**77.8**	**82.1**	69.6
pixel-based methods			
MR Two-Sream R-CNN Peng and Schmid (2016)	71.1	-	-
Attention Pooling Girdhar and Ramanan (2017)	-	52.2	-
Res3D Tran et al. (2017)	-	54.9	85.8
Two-Stream Simonyan and Zisserman (2014)	-	59.4	88.0
IDT Wang and Schmid (2013)	-	61.7	86.4
Dynamic Image Networks Bilen et al. (2016)	-	65.2	89.1
C3D (3 nets) Tran et al. (2015)+IDT	-	-	90.4
LatticeLSTM Sun et al. (2017)	-	66.2	93.5
Two-Stream Fusion Feichtenhofer et al. (2016)+IDT	-	69.2	93.5
TSN Wang et al. (2016)	-	69.4	94.2
Spatio-Temporal ResNet Christoph and Pinz (2016)+IDT	-	70.3	94.6
I3D Carreira and Zisserman (2017)	-	80.7	**98.0**
Spatiotemporal Fusion Zhou et al. (2020)	-	-	96.5
TEA Li et al. (2020)	-	73.3	96.9
2D geometry and pixel combined			
Chained (Pose+RGB+Flow) Zolfaghari et al. (2017)	76.1	69.7	91.1
I3D+PoTion Choutas et al. (2018)	85.5	80.9	98.2
I3D+PA3D Yan et al. (2019)	-	82.1	-
I3D'+SCN (Ours)	**86.3**	**85.1**	**98.3**



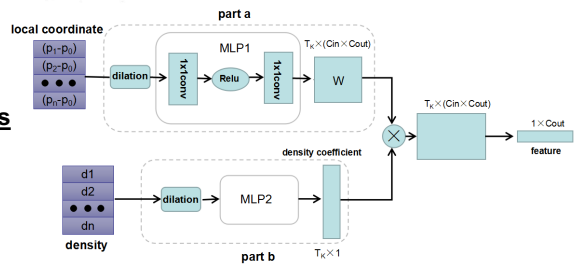
- SCNN outperforms all geometry-based state-of-the-art methods on the three benchmark datasets (JHMDB, HMDB, UCF 101)
- SCNN performs the best on small training datasets compared to all other methods
- SCNN, when integrated with I3D achieves the best accuracy out of all other methods

Framework

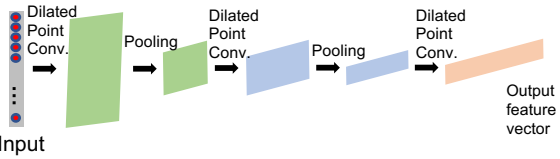
Novel stacked silhouette point representation for video action
Modified Mask-RCNN extracts silhouettes from each frame of the video which are stacked along time to form a 3D point cloud.



Dilated Silhouette Point Convolutions
Performed in a local silhouette patch to obtain a feature output.



Dilated Silhouette Convolutional Neural Network



Slow-to-Fast Network

Conclusions

- SCNN uses the novel geometric representation of silhouettes stacked along the time axis to explore spatio-temporal dynamics.
- SCNN computes dilated silhouette convolutions yielding distinctive geometric features for accurate action recognition.
- SCNN outperforms similar state-of-the-art methods, and when combined with I3D, SCNN achieves the best performance.
- SCNN is implemented in my Coach AI app to increase the accuracy and precision of complex action recognition and the feedback provided to users. I will also further develop my app to assist with physical therapy in the future.