中国科学院软件研究所学术年会’2020
暨计算机科学国家重点实验室开放周

学术论文

Image-based Pose Representation for Action Recognition and Hand Gesture Recognition

IEEE Conference on Automatic Face and Gesture Recognition 2020

基于姿态编码的动作识别与手势识别

林泽一，张维，邓小明，马翠霞，王宏安

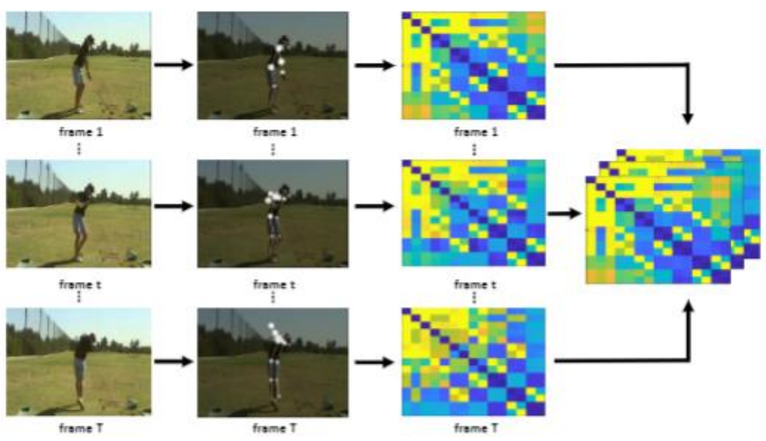
联系人：邓小明　　电话：13717981135　　邮箱：xiaoming@iscas.ac.cn

1. Motivation

- Connection between different human/hand skeleton joints is an important feature in action recognition and hand gesture recognition.

- 3DCNN architecture can extract spatial-temporal feature effectively in image-based action recognition and hand gesture recognition.

2. Poseimage

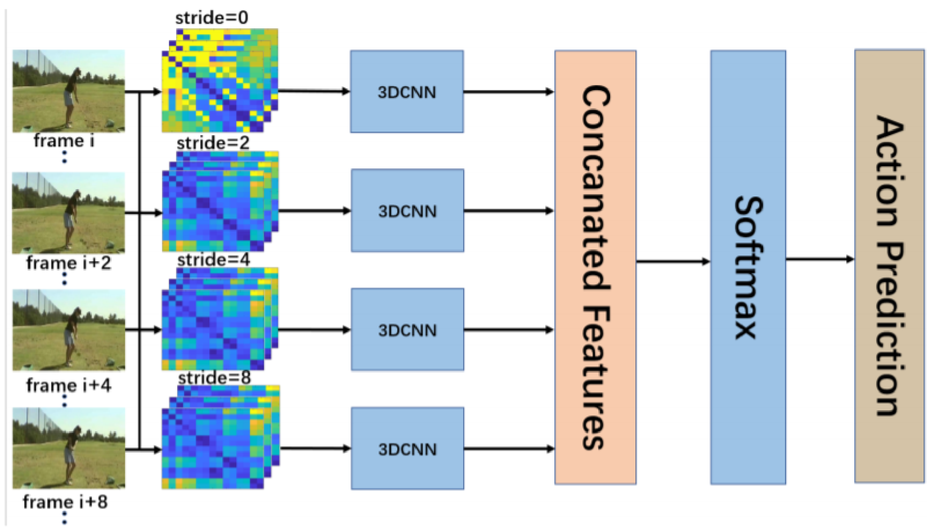


Given a video, we can get the human joints or hand joints $\{J_i\}$of each frame. using the state-of-the-art method OpenPose. Then we calculate the pairwise Euclidean distance $d_{ij} = |J_i - J_j|_2$ and geodesic distance g_{ij} between the joints. We encode the human/hand pose as a Poseimage I of $J \times J$ pixels, where the value at I_{ij} is defined in below and the normalized distance between the i-th and j-th joints.

$$I_{ij} = \frac{d_{ij}}{g_{ij}}$$

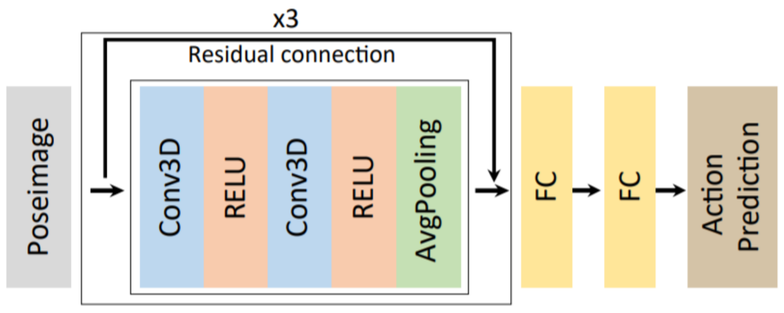
Poseimage has good property of similarity invariance, which could reduce the effect of anthropometry scale changes or viewpoint variations. (See our paper for detailed proof)

3. Poseimage Pyramid



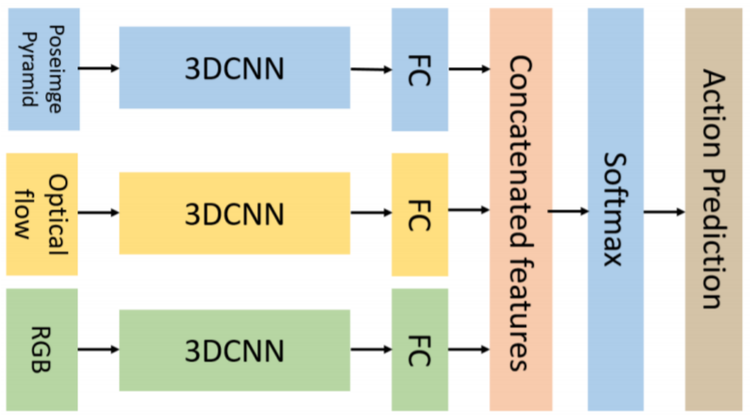
In order to encode the multi-scale temporal information of human/hand pose, we use different temporal strides and get a sequence of Poseimages, which is named as Poseimage Pyramid inspired by image pyramid. Unlike the original Poseimage which calculated by the joints in the same frame, we construct the Poseimage Pyramid with the joints in different frames, and feed Poseimage Pyramid to 3DCNNs.

4.Network Architecture



Our pose stream network architecture. We use Poseimage as input, and use 3 convolution blocks for feature extraction, then use two fully-connected layers and a softmax function to get action or hand gesture recognition result.

5. Integration with I3D



Our multi-stream network architecture for action and hand gesture recognition.

6. Experiments

Our method achieves state-of-the-art performance on the main benchmark dataset such as UCF101,HMDB,JHMDB,NTU-RGBD and SHREC2017 datasets.

Method	JHMDB	HMDB	UCF-101
P-CNN [19]	61.1	-	-
Action Tubes [35]	62.5	-	-
MR Two-Stream RCNN[36]	71.1	-	-
Chained(Pose+RGB+FLOW) [16]	76.1	69.7	91.1
Potion+I3D [17]	85.5	80.9	98.2
Dynamic Image Networks[37]	-	65.2	89.1
C3D(3 nets)+IDT [13]	-	-	90.4
Two-Stream Fusion+IDT [8]	-	69.2	93.5
LatticeLSTM [38]	-	66.2	93.6
TSN [14]	-	69.4	94.2
Spatial-Temporal ResNet+IDT[39]	-	70.3	94.6
I3D [15]	-	80.7	98.0
SVMP+I3D [40]	-	81.3	-
Poseimage Pyramid	76.1	42.3	58.0
I3D	87.8	80.6	98.0
I3D+Poseimage Pyramid	**90.4**	**81.3**	**98.2**

Method	CS	CV
Deep LSTM[42]	60.7	67.3
PA-LSTM[42]	62.9	70.3
ST-LSTM+TS[46]	69.2	77.7
Temporal Conv[43]	74.3	83.1
JDM[1]	76.2	82.3
C-CNN+MTLN[44]	79.6	84.8
ST-GCN[20]	81.5	88.3
3DM[41]	82.0	89.5
DPRL[45]	83.5	89.8
Poseimage Pyramid	**84.0**	**90.5**

Method	14 gestures	28 gestures
De Smedt et al[6]	88.2	81.9
Devineau et al. [7]	91.2	84.3
ST-GCN[20]	92.7	87.7
STA-Res-TCN[4]	93.6	90.7
ST-TS-HGR-NET[5]	94.29	89.4
DG-STA[3]	**94.4**	90.7
Poseimage Pyramid	**94.4**	**91.1**