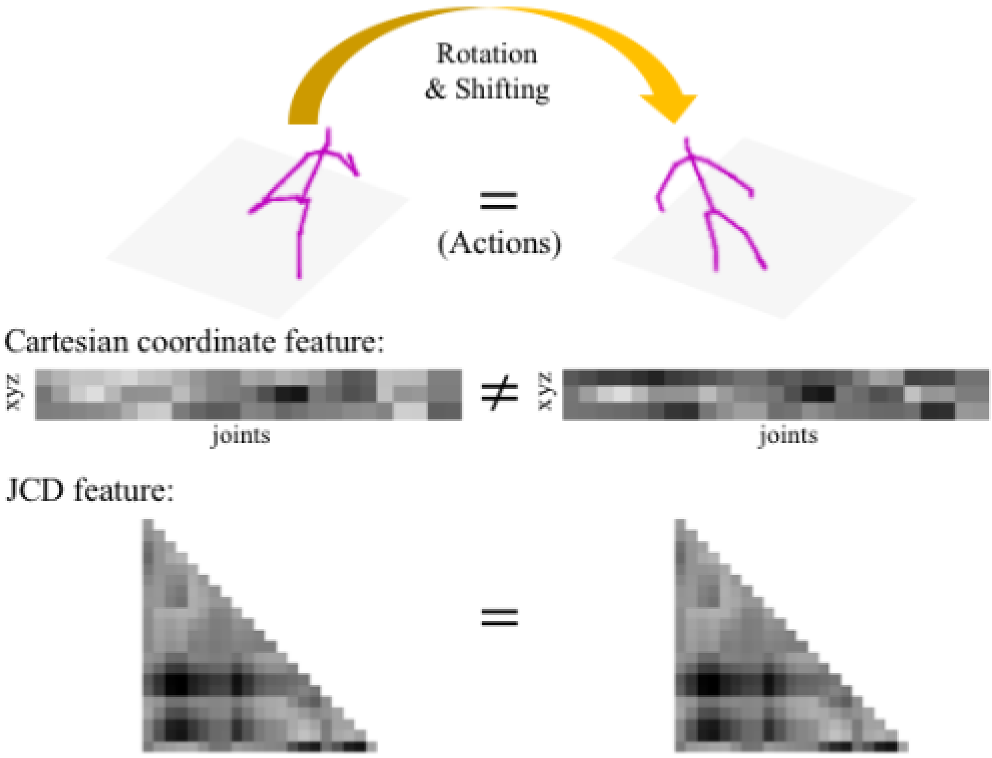
状态

- 已开源
- 全文链接

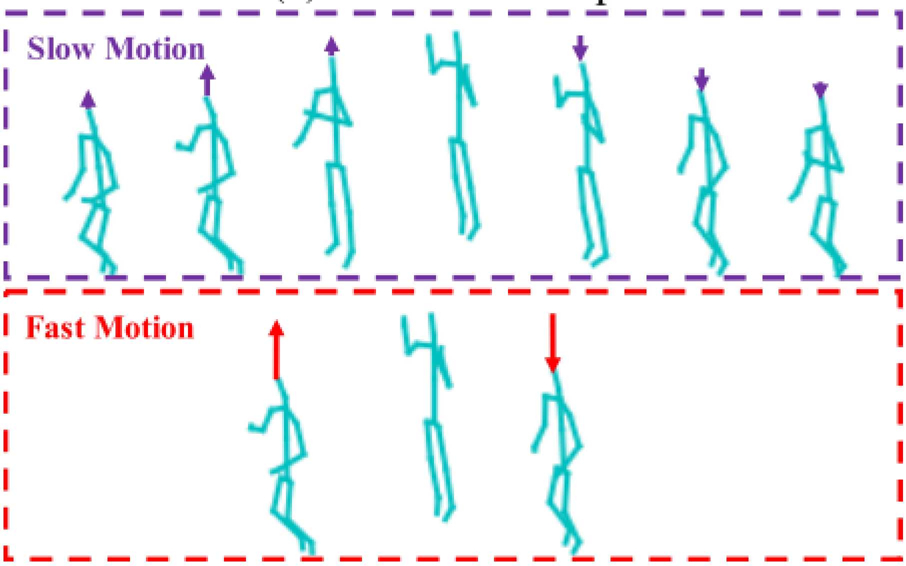
简介

通过骨骼序列做动作识别的一些常见问题：

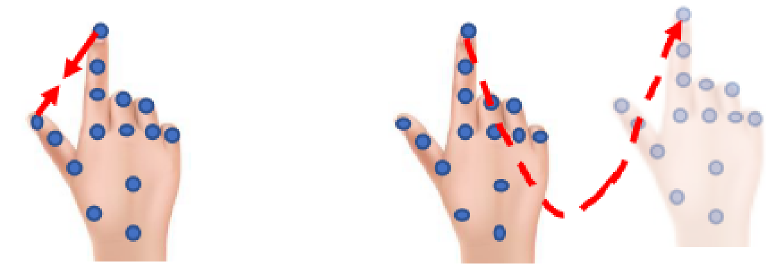
1. 本地视角的变换
2. 动作快慢的不同
3. 动作是否跟全局轨迹相关
4. 不相关的关节索引



(a) Location-viewpoint variation



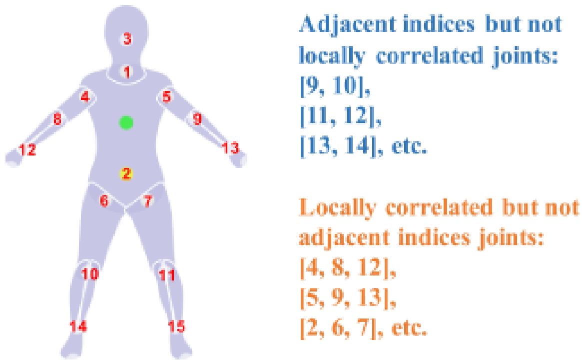
(b) Motion scale variation



Actions unrelated to global trajectories, e.g., pinch.

Actions related to global trajectories, e.g., swipe V.

(c) Related/unrelated to global trajectories



(d) Uncorrelated joint indices (PuppetModel [9])

Fig. 1: Examples of skeleton sequence properties.

方法概述

网络DDNet的结构如下图所示:

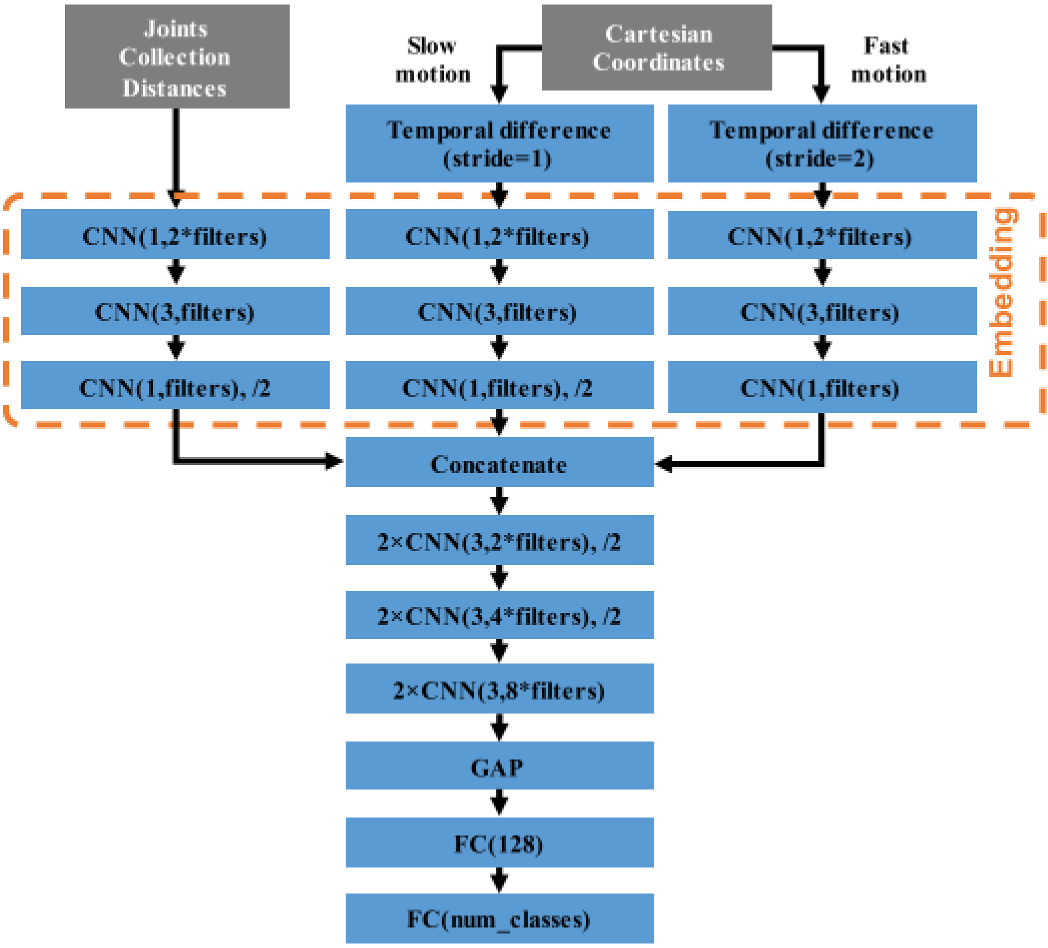


Fig. 2: The network architecture of DD-Net. “2×CNN(3, 2*$filters$), /2” denotes two 1D ConvNet layers (kernel size = 3, channels = 2*$filters$) and a Maxpooling (strides = 2). Other ConvNet layers are defined in the same format. GAP denotes Global Average Pooling. FC denotes Fully Connected Layers (or Dense Layers). We can change the model size by modifying $filters$.

利用Joint Collection Distances提取本地视角不变特征

对于基于骨架的动作识别，一般要么用几何特征、要么直接使用笛卡尔的坐标特征。
一般来讲，对于同样的动作，笛卡尔坐标特征是随着位置和视角变化较大的，而几何特征变化相对较小，但是几何特征需要针对不同的数据集去设计，而且包含的冗余信息也很多。所以作者提出了用JDC feature缓解这些问题。

先计算一对关节坐标的欧式距离，得到哟个对称阵，然后取不含对角元素的下三角阵。

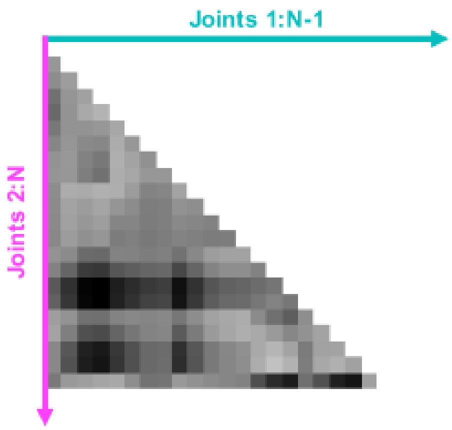


Fig. 3: An example of Joint Collection Distances (JCD) feature at frame k, where the number of joints is N.

$$JCD^k = \begin{bmatrix} \left\| \overrightarrow{J_2^k J_1^k} \right\|_2 & & & \\ \vdots & \ddots & & \\ \vdots & \cdots & \ddots & \\ \left\| \overrightarrow{J_N^k J_1^k} \right\|_2 & \cdots & \cdots & \left\| \overrightarrow{J_N^k J_{N-1}^k} \right\|_2 \end{bmatrix}; \quad (1)$$

where $\left\| \overrightarrow{J_i^k J_j^k} \right\|_2 (i \neq j)$ denotes the Euclidean distance between J_i^k and J_j^k.

利用双尺度特征来提取全局尺度不变动作

JCD feature的缺点是没有考虑到全局轨迹，信息是不充分的。可以用时序信息的差异（比如笛卡尔坐标的速度）来获取全局的动作信息。但是对于同样的动作，其速度尺度可能是不同的。所以作者提出了一个快慢尺度特征：

$$\begin{aligned} M_{slow}^k &= S^{k+1} - S^k; \\ M_{fast}^k &= S^{k+2} - S^k; \end{aligned} \quad (2)$$

通过Embedding来提取关节间的关联

不同的动作利用到的关联特征信息是不同的，跟关节的索引关系并不确定。
为了更好的利用关节间的关联信息，作者将JCD feature和双尺度动作特征做了embedding，将其投

影到隐空间当中。而且可以一定程度上降低噪声的影响：

More formally, let embedding representations of JCD^k, M^k_{slow} and M^k_{fast} to be ε^k_{JCD}, $\varepsilon^k_{M_{slow}}$ and $\varepsilon^k_{M_{fast}}$, respectively, the embedding operation is as follows,

$$\begin{aligned} \varepsilon^k_{JCD} &= Embed_1(JCD^k); \\ \varepsilon^k_{M_{slow}} &= Embed_1(M^k_{slow}); \\ \varepsilon^k_{M_{fast}} &= Embed_2(M^k_{fast}). \end{aligned} \tag{3}$$

where the $Embed_1$ is defined as $Conv1D(1, 2 * filters) \rightarrow Conv1D(3, filters) \rightarrow Conv1D(1, filters)$, and the $Embed_2$ is defined as $Conv1D(1, 2 * filters) \rightarrow Conv1D(3, filters) \rightarrow Conv1D(1, filters) \rightarrow Maxpooling(2)$, because JCD^k and M^k_{slow} have double the temporal length of M^k_{fast}.

DD-Net futher concatenates embedding features to a representation ε^k by

$$\begin{aligned} \varepsilon^k &= \varepsilon^k_{JCD} \oplus \varepsilon^k_{M_{slow}} \oplus \varepsilon^k_{M_{fast}}, \\ & w.r.t.\ \ \varepsilon^k \in \mathbb{R}^{(K/2)\times filters}; \end{aligned} \tag{4}$$

where $\oplus$ is the concatenation operation.

After the embedding process, subsequent processes are not affected by the joint indices, and therefore DD-Net can use the 1D ConvNet to learn the temporal information as Fig. 2 shows.

实验

实验数据用的SHREC和JHMDB

TABLE II: Results on SHREC (Using 3D skeletons only) [4]

Methods	**Parameters**	**14 Gestures**	**28 Gestures**
Dynamic hand [19] (CVPRW16)	N/A	88.2%	81.9%
Key-frame CNN [4] (3DOR17)	7.92 M	82.9%	71.9%
3 Cent [21] (STAG17)	N/A	77.9%	N/A
Parallel CNN [5] (RFIAP18)	13.83 M	91.3%	84.4%
STA-Res-TCN [6] (Gesture18)	5-6 M	93.6%	90.7%
MFA-Net [23] (Sensor19)	N/A	91.3%	86.6%
DD-Net (filters=64, w/o global fast&slow motion)	1.70 M	55.2%	41.6%
DD-Net (filters=64, w/o global slow motion)	1.76 M	92.7%	90.2%
DD-Net (filters=64, w/o global fast motion)	1.76 M	93.3%	90.5%
DD-Net (filters=64)	1.82 M	**94.6%**	**91.9%**
DD-Net (filters=32)	0.50 M	93.5%	90.4%
DD-Net (filters=16)	**0.15 M**	91.8%	90.0%

TABLE III: Results on JHMDB (Using 2D skeletons only) [9]

Methods	**Parameters**	**Manually annotated skeletons**
Chained Net [7] (ICCV17)	17.50 M	56.8%
PoTion [8] (CVPR18)	4.87 M	62.1%
EHPI [28] (arXiv19)	1.22 M	65.5%
DD-Net (filters=32, w/o global fast&slow motion)	0.46 M	71.4%
DD-Net (filters=32, w/o global slow motion)	0.48 M	74.9%
DD-Net (filters=32, w/o global fast motion)	0.48 M	75.8%
DD-Net (filters=32)	0.50 M	**78.0%**
DD-Net (filters=64)	1.82 M	77.8%
DD-Net (filters=16)	**0.15 M**	74.7%

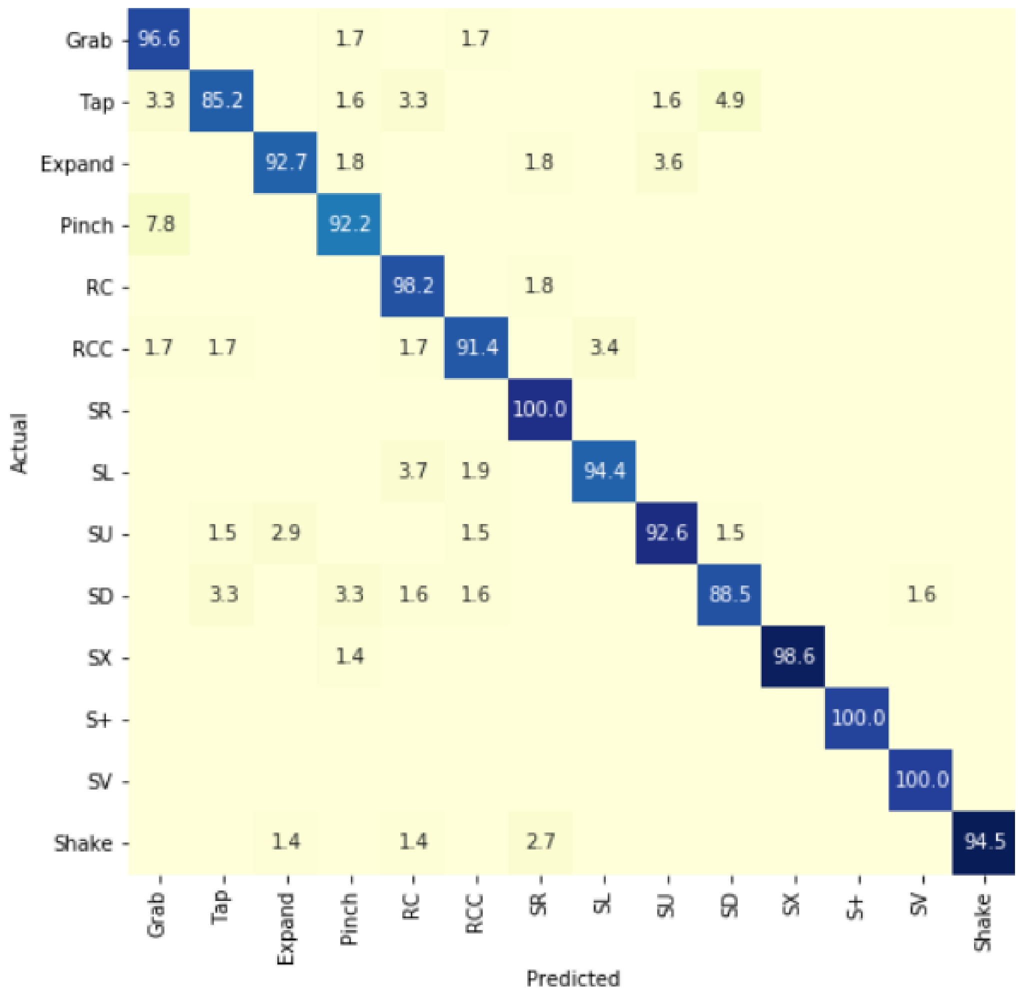


Fig. 4: Confusion matrix of SHREC dataset (14 hand actions) obtained by DD-Net.

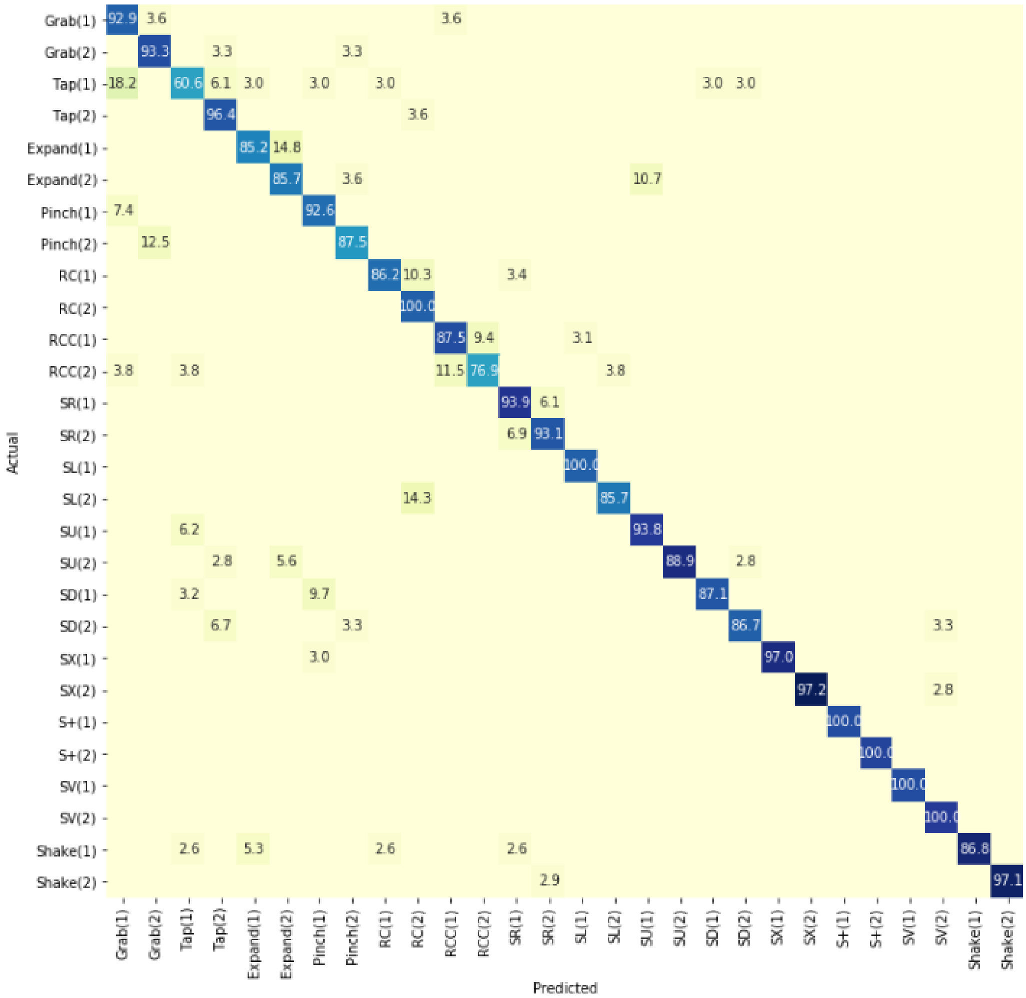


Fig. 5: Confusion matrix of SHREC dataset (28 hand actions) obtained by DD-Net.

通过实验可以发现，每个模块都有作用而且模型的速度非常快

总结

对输入进行预处理，加入一些trick使得模型学习的降低，其实是简化模型结构的一个非常有效方法，但是这个比较局限于输入不是单纯RGB图像的时候。同时作者也说了结合RGB图像和深度信息能够做到更高的精度。