行为识别：让机器学会“察言观色”第一步

电影短片《Changing Batteries》讲了这样一个故事：独居的老奶奶收到儿子寄来的一个机器人，这机器人善于察言观色，很快就跟老奶奶“心有灵犀”，不仅能在老奶奶口渴时为她端水、在老奶奶扫地时接过老奶奶的扫把，做力所能及的家务活，如果老奶奶在椅子上看电视睡着了，机器人还为她轻轻盖上踏足。有了它，老奶奶又重新感受到久违的快乐，过上了更轻松的生活[1]……咳咳，催泪的故事讲完了，接下来我们先说说这机器人的察言观色技能是怎么实现的。

在人工智能研究领域，这一技能叫人体行为识别，是智能监控、人机交互、机器人等诸多应用的一项基础技术。以电影提到的老人智能看护场景为例，智能系统通过实时检测和分析老人的行动，判断老人是否正常吃饭、服药、是否保持最低的运动量、是否有异常行动出现（例如摔倒），从而及时给予提醒，确保老人的生活质量不会由于独自居住而有所降低。第二个例子是人机交互系统，通过对人的行为进行识别，猜测用户的“心思”，预测用户的意图，及时给予准确的响应。第三个例子是医院的康复训练，通过对动作行为的规范程度做出识别，评估恢复程度以提供更好的康复指导等。

俗话说“排骨好吃，骨头难啃”，行为识别是一项具有挑战性的任务，受光照条件各异、视角多样性诸多因素的影响。对行为识别的研究可以追溯到1973年，当时Johansson通过实验观察发现，人体节点的移动来描述，因此，只要10-12个关键节点的组合与追踪便能形成对诸多行为例如跳舞、走路人体关键节点的运动来识别行为[2]。正因为如此，在Kinect的游戏中，系统根据深度图估计出的人

的一些关节点的位置信息组成），对人的姿态动作进行判断，促成人机交互的实现。另一个重要分支则是基于RGB视频做行为动作识别。与RGB信息相比，骨架信息具有特征明确简单、不易受外观因素影响的优点。我们在这里主要探讨基于骨架的行为识别及检测。

人体骨架怎么获得呢？主要有两个途径：通过RGB图像进行关节点估计（Pose Estimation）获得[3][4]，或是通过深度摄像机直接获得（例如Kinect）。每一时刻（帧）骨架对应人体的K个关节点所在的坐标位置信息，一个时间序列由若干帧组成。行为识别就是对时域预先分割好的序列判定其所属行为动作的类型，即“读懂行为”。但在现实应用中更容易遇到的情况是序列尚未在时域分割（Untrimmed），因此需要同时对行为动作进行时域定位（分割）和类型判定，这类任务一般称为行为检测。

基于骨架的行为识别技术，其关键在于两个方面：一方面是如何设计鲁棒和有强判别性的特征，另一方面是如何利用时域相关性来对行为动作的动态变化进行建模。

我们采用基于LSTM （Long-Short Term Memory）的循环神经网络（RNN）来搭建基础框架，用于学习有效的特征并且对时域的动态过程建模，实现端到端（End-to-End）的行为识别及检测。关于LSTM的详细介绍可参考[5]。我们的工作主要从以下三个方面进行探讨和研究：

- 如何利用空间注意力（Spatial Attention）和时间注意力（Temporal Attention）来实现高性能行为动作识别 [8]？
- 如何利用人类行为动作具有的共现性（Co-occurrence）来提升行为识别的性能[7]？
- 如何利用RNN网络对未分割序列进行行为检测（行为动作的起止点的定位和行为动作类型的判定）[9]？

空时注意力模型（Attention）之于行为识别

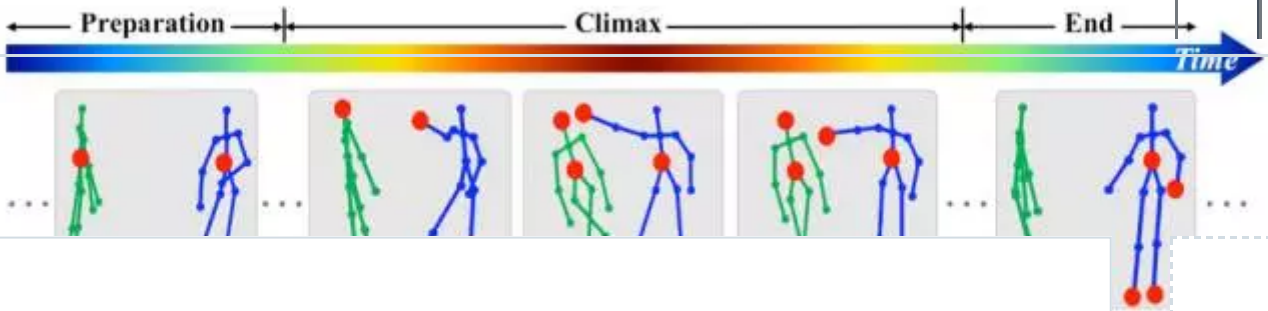


图1.1：“挥拳”行为动作序列示例。行为动作要经历不同的阶段（比如靠近、高潮、结束），涉及到不同的具有判别力的关节点子集合（如红色圆圈所示）。这个例子中，人体骨架由15个关节点的坐标位置表示。

注意力模型（Attention Model）在过去这两年里成了机器学习界的“网红”，其想法就是模拟人类对事物的认知，将更多的注意力放在信息量更大的部分。我们也将注意力模型引入了行为识别的任务，下面就来看一下注意力模型是如何在行为识别中大显身手的。

时域注意力：众所周知，一个行为动作的过程要经历多个状态（对应很多时间帧），人体在每个时刻也呈现出不同的姿态，那么，是不是每一帧在动作判别中的重要性都相同呢？以“挥拳”为例，整个过程经历了开始的靠近阶段，高潮阶段及结束阶段。相比之下，挥动拳脚的高潮阶段包含了更多的信息，最有助于动作的判别。依据这一模型，通过一个LSTM子网络来自动学习和获知序列中不同帧的重要性，使重要的帧在分类中起更大的作用。

空域注意力：对于行为动作的判别，是不是每个关节点在动作判别中都同等重要呢？研究证明，一些行为动作会跟某些关节点构成的集合相关，而另一些行为动作会跟其它一些关节点构成的集合相关。比如“打电话”，主要跟头、肩膀、手肘和手腕这些关节点密切相关，同时跟腿上的关节点关系很小，而对“走路”这个动作的判别主要通过腿部节点的观察就可以完成。与此相适应，我们设计了一个LSTM子网络，依据序列的内容自动给不同关节点分配不同的重要性，即给予不同的注意力。由于注意力是基于内容的，即当前帧信息和历史信息共同决定的，因此，在同一个序列中，关节点重要性的分配可以随着时间的变化而改

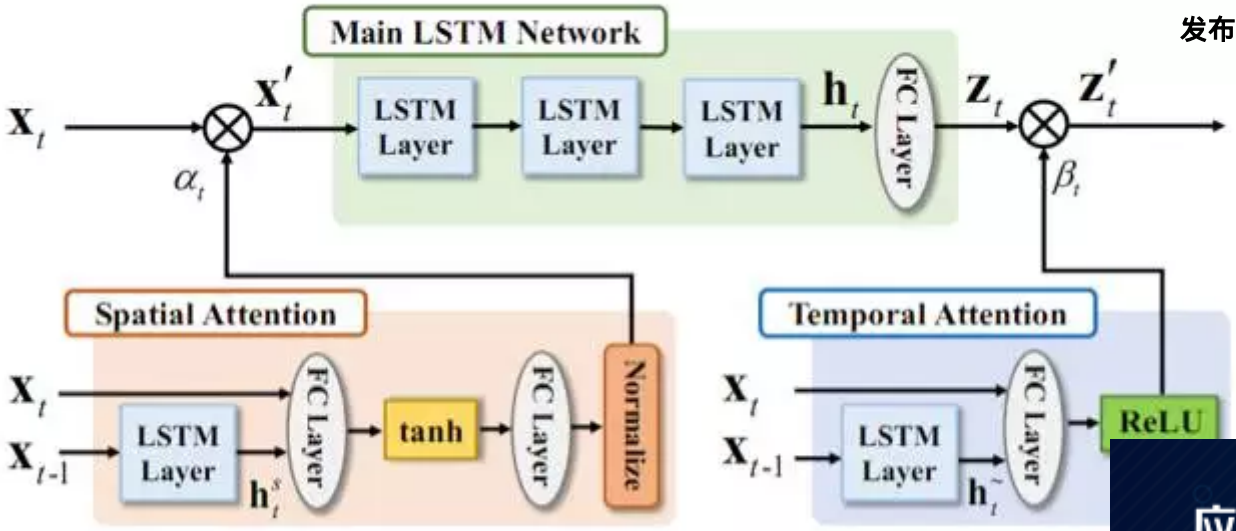


图1.2：网络结构框图。主网络（Main LSTM Network）用于对特征进行提取、时域相关性利用和最
络（Temporal Attention）用于给不同帧分配合适的重要性。空域注意力子网络（Spatial Attention
适的重要性。

空时注意力模型能带来多大的好处呢？我们在SBU 数据库、NTU RGB+D 数据库的Cross Subject(CS) 和 Cross View(CV) 设置上分别进行了实验，以检测其有效性。图1.3展示了性能的比较：LSTM表示只有主LSTM网络时的性能（没引入注意力模型）。当同时引入时域注意力（TA）和空域注意力（SA）网络后，如STA-LSTM所示，识别的精度实现了大幅提升。

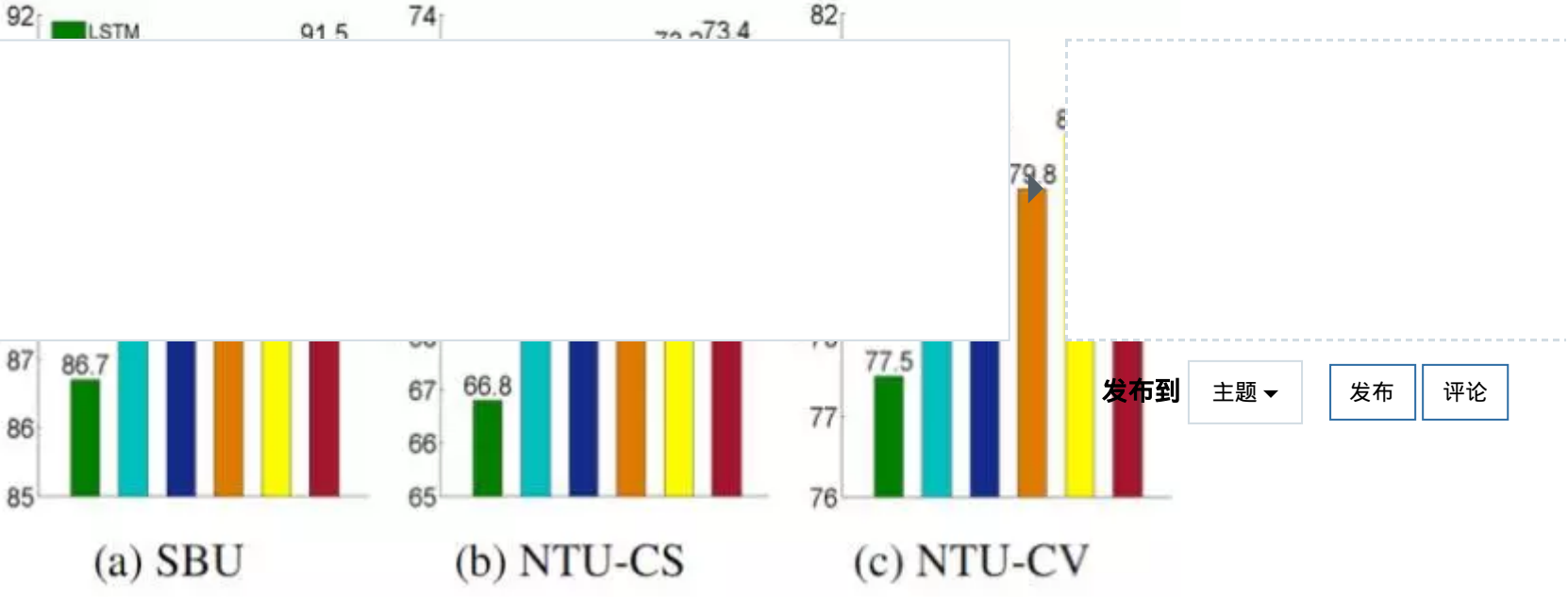


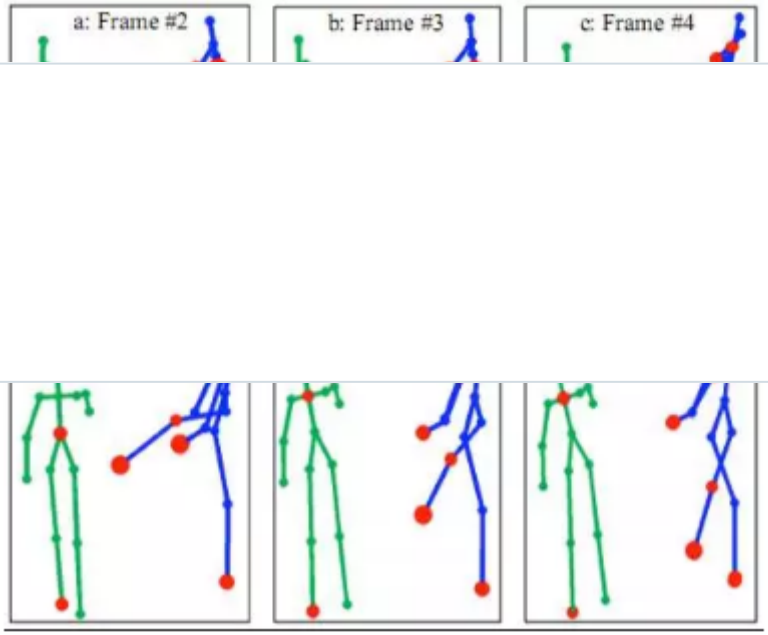
图1.3：空时注意力网络的识别精度比较。(a) SBU 数据库。(b) NTU 数据库Cross Subject(CS)。(c) NTU数据库Cross-View(CV)。其中, LSTM只包含主网络结构。STA-LSTM同时包含了空时子网络。

细心的读者可能已经发现，序列中的空域注意力和时域注意力具体为多大是没有参考的(不知道Gr
终分类性能来自动习得注意力。那么，学到的注意力模型分配的注意力数值是什么样呢？我们可
的输出。图1.4可视化了在 “挥拳”行为动作的测试序列上，模型输出的空域注意力权重的大小，时
域注意力的差值。如图1.4（a）中所示，主动方（右侧人）的节点被赋予了更大的权值，且腿部的

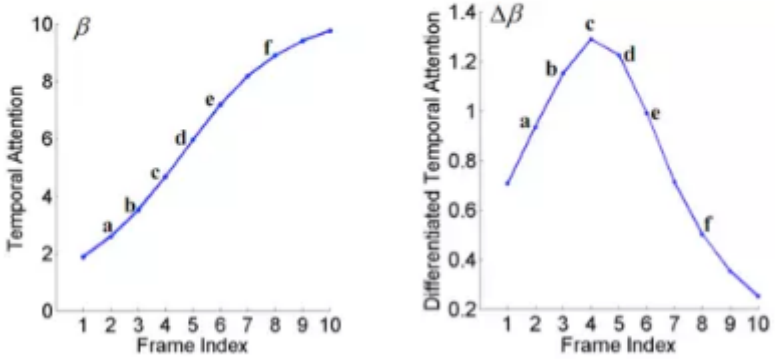
(http://www.csdn.net/?ref=toolbar)

增量。时域注意力模型会对更具判别力的帧赋予较大的注意力权重。对不同的行为动作，空间注意力模型赋予较大权重的节点也不同，整体和人的感知一致。

请输入标题
请输入链接地址



(a)



(b) (c)

请输入推荐理由

请输入标签

发布到 主题 发布 评论

关闭

图1.4. 空时注意力模型学到的权重在“挥拳”测试序列上的可视化。(a) 空域注意力权重。红色圆圈的大小示意对应关节点权重的大小。红色圆圈越大，表示权重越大。这里我们只将有着最大权重的前8个节点做了标记。(b) 时域注意力权重。(c) 差分时域注意力权重，即相邻帧的时域注意力权重的差值。

LSTM网络框架和关节点共现性（Co-occurrence）的挖掘之于行为识别

同。例如对于“走路”的行为动作，“脚腕”、“膝盖”、“臀部”等关节点构成具有判别力的节点集合。我们将这种几个关节点同时影响和决定判别的特性称为共现性（Co-occurrence）。

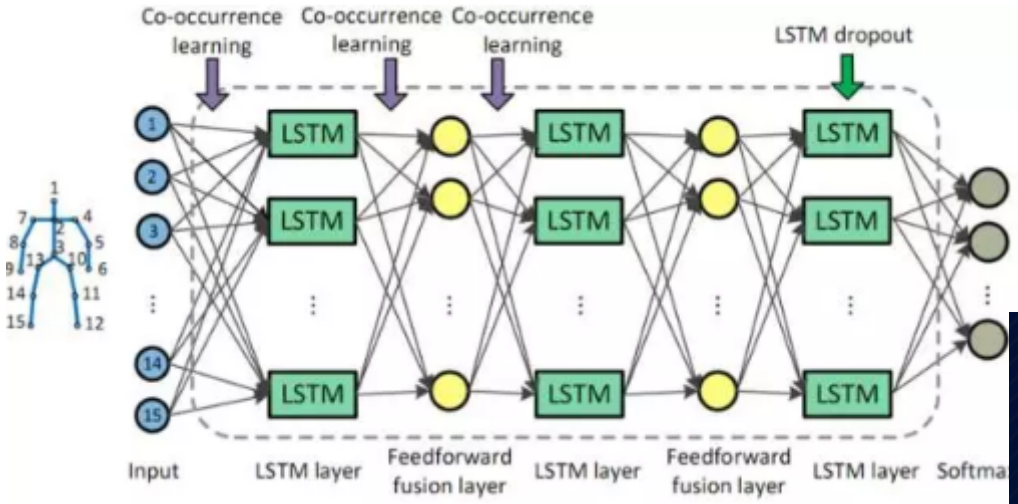


图 2.1 基于LSTM的网络结构和共现性特性的利用。

在训练阶段，我们在目标函数中引入对关节点和神经元相连的权重的约束，使同一组的神经元对某些关节点组成的子集有更大的权重连接，而对其他节点有较小的权重连接，从而挖掘关节点的共现性。如图2.2所示，一个LSTM 层由若干个LSTM神经元组成，这些神经元被分为K组。同组中的每个神经元共同地和某些关节点有更大的连接权值（和某类或某几类动作相关的节点组成关节点子集），而和其他关节点有较小的连接权值。不同组的神经元对不同动作的敏感程度不同，体现在不同组的神经元对应于更大连接权值的节点子集也不同。在实现上，我们通过对每组神经元和关节点的连接加入组稀疏（Group Sparse）约束

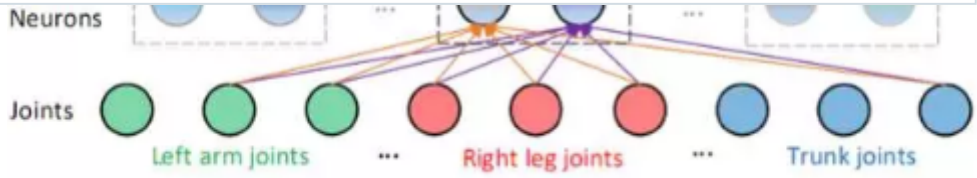
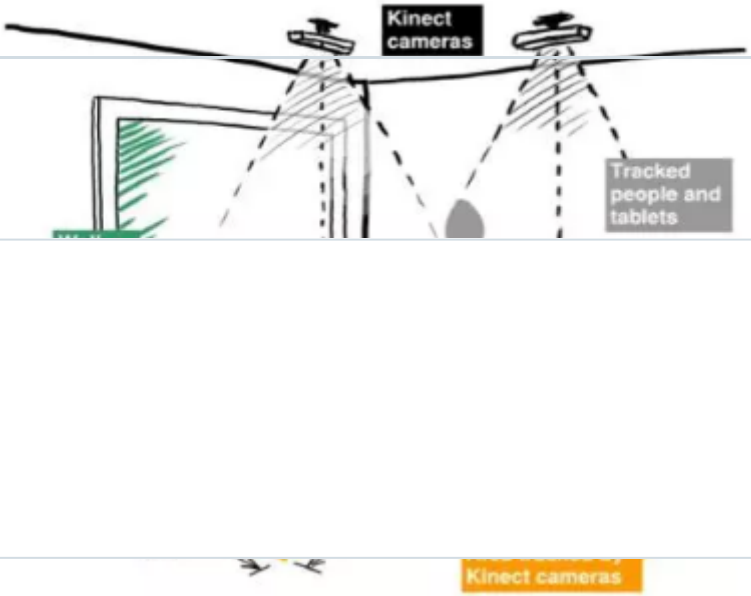


图2.2 第一层的神经元(LSTM Neurons)和关节点连接的示意图。以第k组的神经元为例，第k组的神经元都同时对某几个关节点有着大的权重连接，而对其他关节点有着小的权重连接（在这里用未连接来示意）。

基于联合分类和回归的循环神经网络之于行为动作检测



（图片来自网络）

前面讨论了对于时域分割好的序列的行为动作分类问题。但是想要计算机get到“察言观色”的技能并不那么容易。在实际的应用中多有实时的需求，而摄像头实时获取的视频序列并没有根据行为动作的发生位置进行预先时域分割，因此识别系统不仅需要判断行为动作的类型，也需要定位行为动作发生的位置，即进行行为动作检测。如图3.1所示，对于时间序列流，检测系统在每个时刻给出是否当前是行为动作的开始或结束，以及行为动作的类型信息。

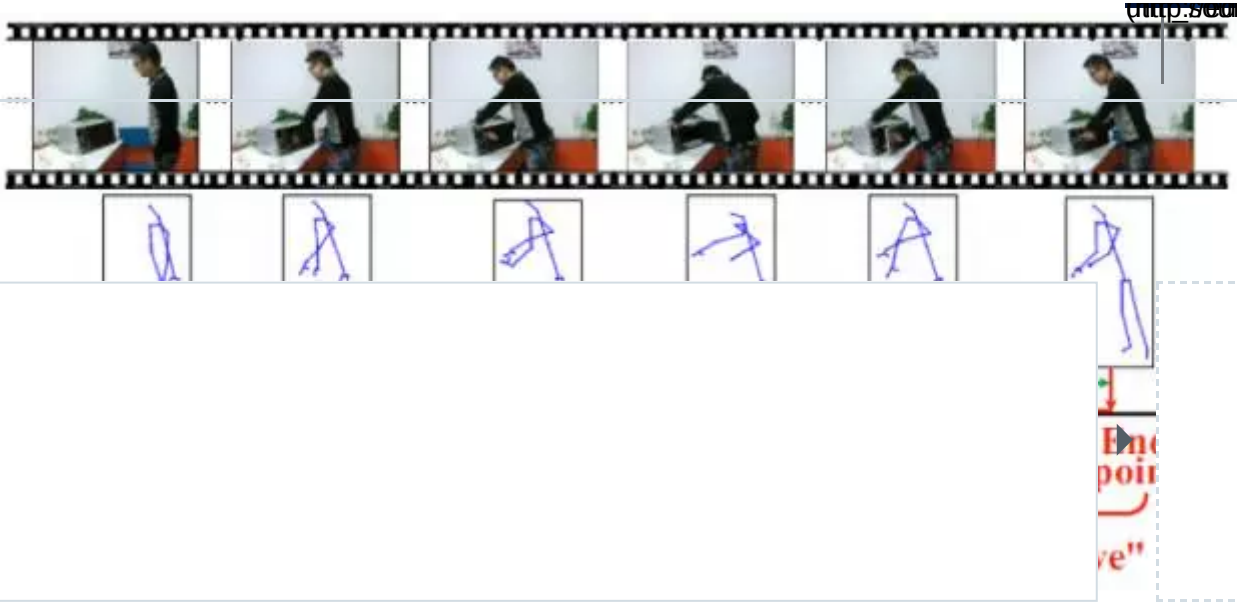


图3.1：行为动作检测示例。对于时间序列流，系统在每个时刻给出是否当前是行为动作的开始或结束，以及行为动作的类型信息。

图3.2：基于滑动窗口的行为动作检测示意图，即每个时刻对固定或者可变的时域窗口内

在线（Online）的行为动作检测常常采用滑窗的方法，即对视频序列流每次观察一个时间窗口内的内容，对其进行分类。然而，基于滑窗的方法常常伴随着冗余的计算，性能也会受到滑动窗口大小的影响。

针对输入的视频序列流，我们设计了基于循环神经网络LSTM的在线行为动作检测系统，在每帧给出行为动作判定的结果。LSTM的记忆性可以避免显式的滑动窗口设计。如图3.3所示，网络由LSTM 层和全连层（FC Layer）组成前端的网络Deep LSTM Network，后面连接的分类网络（Classification Network）用于判定每帧的动作类别，同时，回归网络（Regression Network）

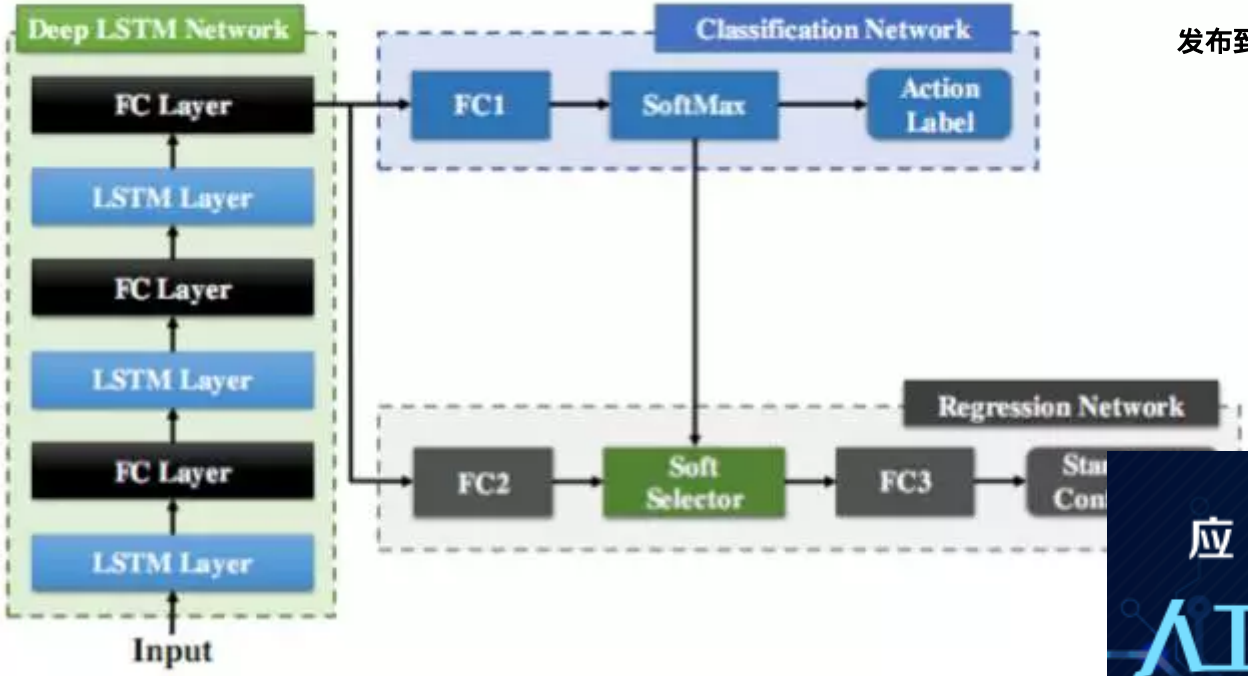


图3.3：用于在线行为动作检测的联合分类回归（Joint Classification-regression）循环网络框架。

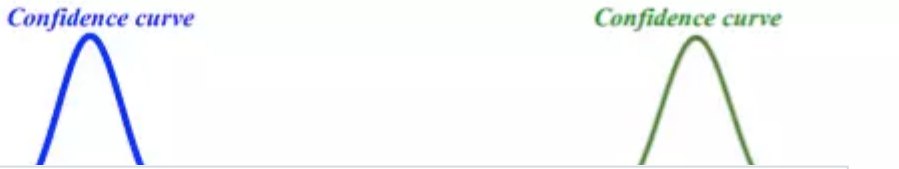


图3.4：行为动作的起止点目标回归曲线。在测试阶段，当起始点（终止点）的回归曲线到达局部峰值时，可以定位为行为动作的起始（结束）位置。

总结和展望

由于行为识别技术在智能监控、人机交互、视频序列理解、医疗健康等众多领域扮演着越来越重要的角色，研究人员正使出“洪荒之力”提高行为识别技术的准确度。说不定在不久的某一天，你家门口真会出现一个能读懂你的行为、和你“心有灵犀”的机器人，对于这一幕，你是不是和我们一样充满期待？

[1] https://movie.douban.com/subject/25757903/ (https://movie.douban.com/subject/25757903/)

[2] Gunnar Johansson. Visual perception of biological motion and a model for it is analysis. Perception and Psychophysics 14(2), pp 201–211, 1973.

[3] Alejandro Newell, Kaiyu Yang, Jia Deng. Stacked Hourglass Networks for Human Pose Estimation, In ECCV, 2016.

[4] Zhe Cao, Tomas Simon, Shih-En Wei, Yaser Sheikh. Realtime Multi-person 2D Pose Estimation using Part Affinity Fields. arXiv

[8] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jiaying Liu. An End-to-End Spatio-Temporal Attention Model for Human Action Recognition from Skeleton Data. Accepted by AAAI, 2017.

[9] Yanghao Li, Cuiling Lan, Junliang Xing, Wenjun Zeng, Chunfeng Yuan, Jiaying Liu. Online Human Action Detection Using Joint Classification-Regression Recurrent Neural Networks. In ECCV, 2016.

作者简介：兰翠玲博士，微软亚洲研究院副研究员，从事计算机视觉，信号处理方面的研究。她的研究兴趣包括行为识别、姿态估计、深度学习、视频分析、视频压缩和通信等，并在多个顶级会议，期刊上发表了近20篇论文，如AAAI, ECCV, TCSVT等。

来源：：微软研究院AI头条，授权CSDN发布。

欢迎人工智能领域技术投稿、约稿、给文章纠错，请发送邮件至heyc@csdn.net (mailto:heyc@c

(http://geek.csdn.net/user/publishlist/heyc861221)
何永灿CSDN (http://geek.csdn.net/user/publishlist/heyc861221)

发布于 人工智能 (http://geek.csdn.net/forum/43) 2017-02-15 10:58

(http://www.csdn.net?ref=toolbar)

请输入标题
请输入链接地址

:)

(http:

请输入推荐理由

请输入标签

发布到 主题 发布 评论

关闭