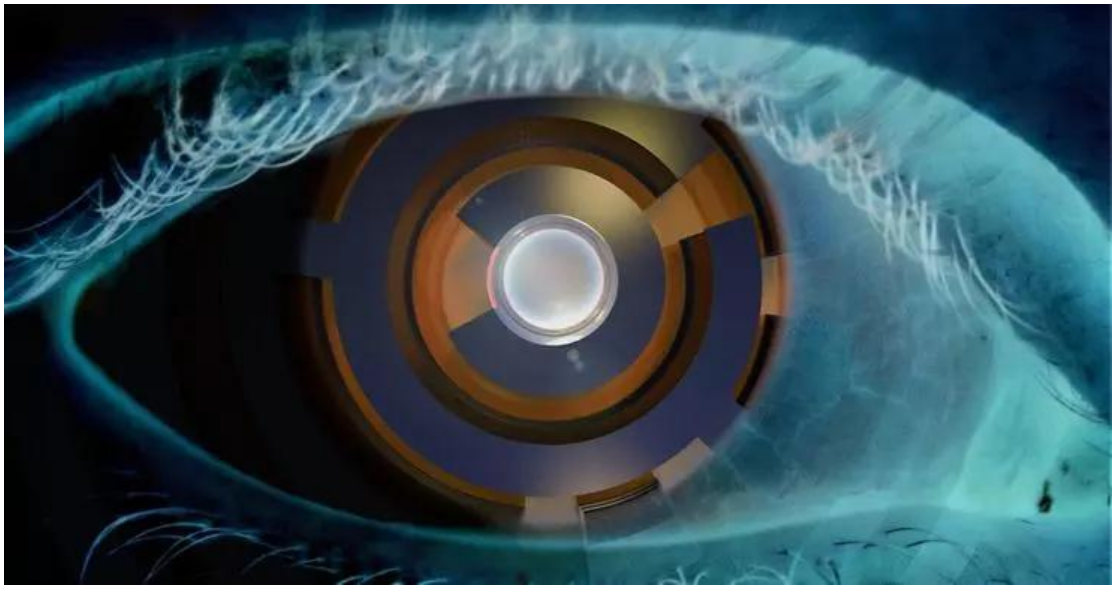
防火防盗防 AI 系列：你的心事，将被你的行走步态暴露！

人工智能7 月 7 日

在未来的世界里，机器人无需与我们产生交流，也能看透我们内心的小九九，这是否听起来有点像是天方夜谭？近期，一支由查珀尔希尔大学（University of Chapel Hill）和马里兰大学（University of Maryland）组成的研究团队，正试图让这一切成为现实。



除了语言，机器还能如何读懂人类的情绪？

情绪毫无疑问在生活中扮演着重要的角色，我们都是通过看别人「脸色」，进而决定下一步采取的应对行为。比如正在生气的女朋友，以及心情大好的女朋友，

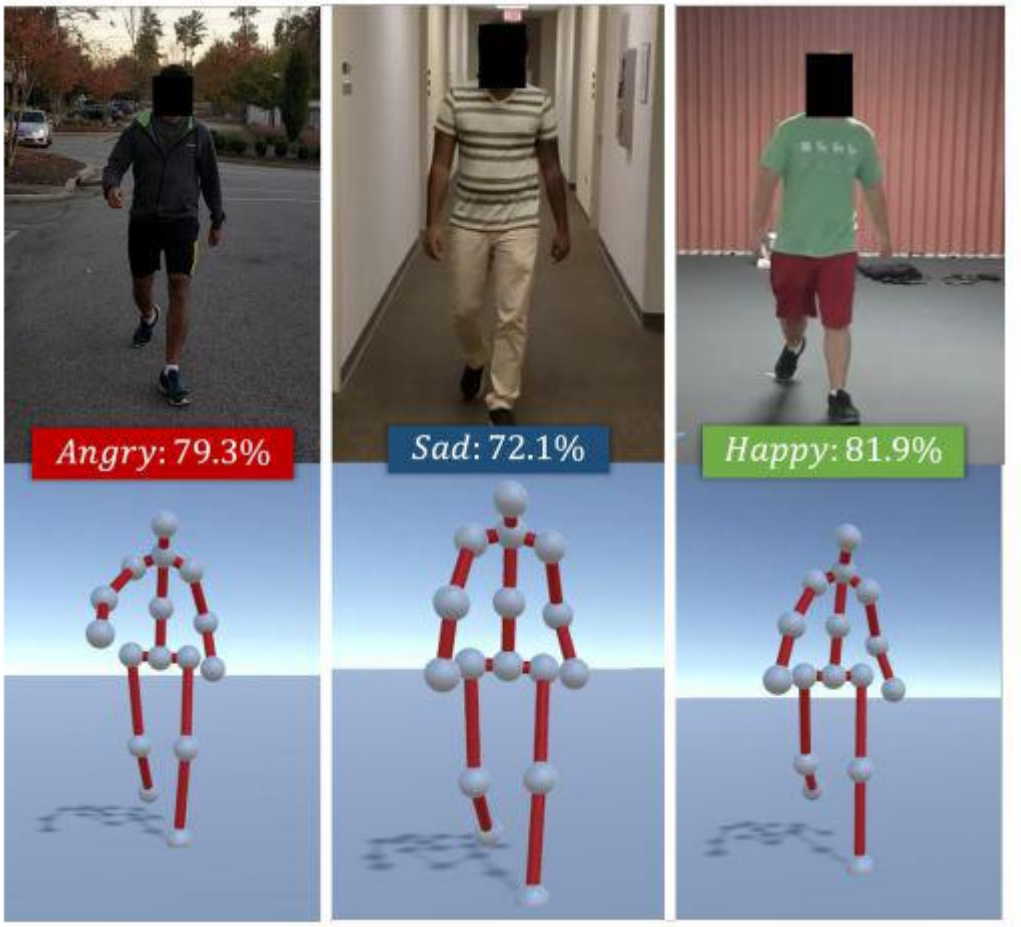
交流使用的肯定不是同一套话术。反过来，很多时候我们也会被他人的情绪影响我们的行为。

因此，自动情绪识别技术是诸多领域的刚需，如游戏娱乐、安保执法、购物、人机交互等。有了它，机器人将能更好地与人类产生交流。对于具备自然语言处理能力的机器人而言，它们可以通过文字/语言交流去推断出用户的情绪，因而问题不大；对于那些不具备相关能力的机器人来说，是否能够通过非语言的方式，比如面部表情或动作姿态，去判断人类当下的情绪状态，依然是一个棘手的问题，目前学界有不少团队正试图为此找到理想方案。

在过去，研究更多集中在帮助机器解读人类丰富表情的含义，然而近期的一些心理学文献却对此提出了质疑——很多种情况下，由于存在一些干扰，人类面部表情不一定代表着对应的交际目的。与此同时，越来越多研究表明，人体行为在情绪传递方面同样扮演者非常重要的角色，而人们在行走时的身体表情或者步态，已经被证明有助于感知情绪。打个比方，当我们沮丧时，上半身会处于耷拉状态，肢体活动速度变慢；当我们快乐时，肢体活动节奏会明显变快，手臂的摆动次数变多。

一个解决方案

在这篇名为《Identifying Emotions from Walking Using Affective and Deep Features》的论文中，研究团队提出了一种全新的自动情绪识别方法，可以将视频中行走的人类进行归类为快乐、悲伤、愤怒或中立 4 种情感类别。



简单来说 ,他们先将这些成功提取出的步态转换为三维形态 ,然后使用基于 LSTM 的方法对这些连贯性的 3D 人体姿势进行长期依赖性建模，以获得深度特征。接着，他们提出了表示人类行走姿势与运动的时空情感身体特征（spatio temporal affective body features），最后将两者进行集合，并使用随机森林分类器（Random Forest Classifier）将成果归类成上述提及的 4 种情感类别。

往细了讲，即是先通过多个步态数据集提取出情感特征——这些情感特征建立在心理表征基础上，当中包括了体态特征和动作特征。接着，通过训练 LSTM 网络进行深度特征提取，然后将深度特征与情感特征相结合，对随机森林分类器进行训练。最后，只要给出一个人行走的 RGB 视频，该 3D 人体步态评估技术将会以 3D 形式对他/她的步态进行解析，进而提取出情感与深层特征，最后再用已经训练好的随机森林分类器来识别出个体的情感状态。

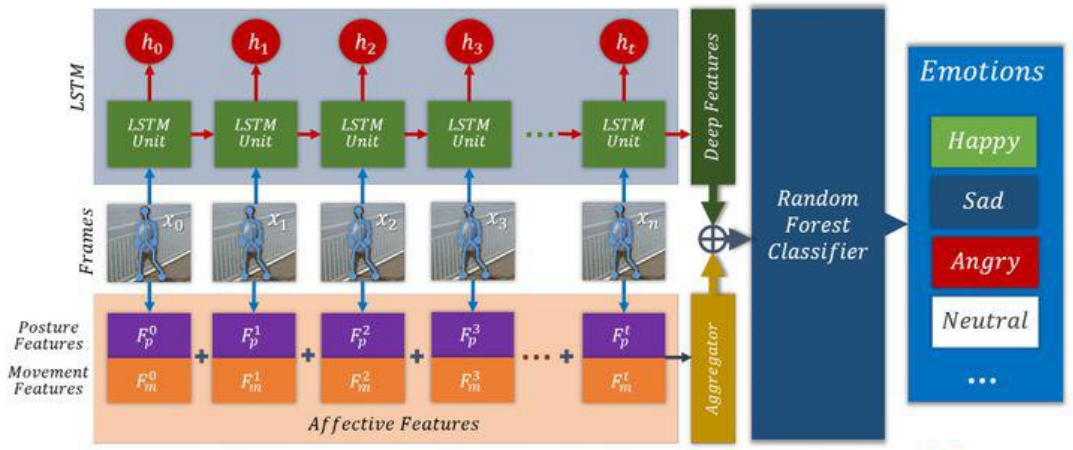


Fig. 3. **Overview:** Given an RGB video of an individual walking, we use a state-of-the-art 3D human pose estimation technique [29] to extract a set of 3D poses. These 3D poses are passed to an LSTM network to extract deep features. We train this LSTM network using multiple gait datasets. We also compute affective features consisting of both posture and movement features using psychological characterization. We concatenate these affective features with deep features and classify the combined features into 4 basic emotions using a Random Forest classifier.

读懂人类情绪的奥秘

要准确评估一个人的情感状态，姿势与运动特征都是必不可少的，其中就包括关节角度、摆动距离、摆动速度以及身体所占空间等特征，都可以被用于识别步态中传递的情感状态。基于这些心理学发现，该团队的工作便将姿势与运动特征都包含了进来。

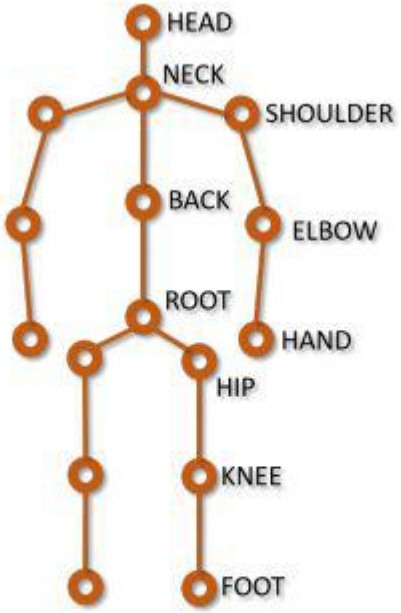


Fig. 4. **Human Representation**: We represent a human by a set of 16 joints. The overall configuration of the human is defined using these joint positions and is used to extract the features.

在姿势特征方面，该团队主要从这几个方面进行了定义：

体积：身体的舒展一般传达的是正面情绪；当一个人在表达负面情绪的时候，身体姿势往往更紧凑。

面积：通过手和颈部之间以及脚和根关节之间的三角区域来模拟身体的扩张情况。

距离：脚和手之间的距离也可用于模拟身体的扩张情况。

角度：头部倾斜情况，通过颈部不同关节延伸的角度来区分快乐和悲伤情绪。

此外，他们还将步幅作为姿势的特征之一——长步幅表示愤怒和快乐；短步幅表示悲伤和中立。

TABLE 1
Posture Features: We extract posture features from an input gait using emotion characterization in visual perception and psychology literature [23], [25].

Type	Description
Volume	Bounding box
Angle	At neck by shoulders
	At right shoulder by neck and left shoulder
	At left shoulder by neck and right shoulder
	At neck by vertical and back
	At neck by head and back
Distance	Between right hand and the root joint
	Between left hand and the root joint
	Between right foot and the root joint
	Between left foot and the root joint
	Between consecutive foot strikes (stride length)
Area	Triangle between hands and neck
	Triangle between feet and the root joint

在运动特征方面，他们则做出以下定义：

与低唤醒情绪相比，高唤醒情绪的运动明显在频次上会更密集。

快步态代表快乐或愤怒；慢步态代表悲伤。

TABLE 2
Movement Features: We extract movement features from an input gait using emotion characterization in visual perception and psychology literature [23], [25].

Type	Description
Speed	Right hand
	Left hand
	Head
	Right foot
	Left foot
Acceleration Magnitude	Right hand
	Left hand
	Head
	Right foot
	Left foot
Movement Jerk	Right hand
	Left hand
	Head
	Right foot
	Left foot
Time	One gait cycle

最终实验结果显示 ,该团队的方案相较其他分类方法 ,准确率更高 ,达到 80:07%；

即便用于非动作数据集（ non-acted data ）上，准确率也高达 79:72%。

TABLE 3
Performance of Different Classification Methods: We analyze different classification algorithms to classify the concatenated deep and affective features. We observe an accuracy of 80.07% with the Random Forest classifier.

Algorithm (Deep + Affective Features)	Accuracy
LSTM + Support Vector Machines (SVM)	70.04%
LSTM + Stochastic Gradient Descent (SGD)	71.01%
LSTM + Random Forest	**80.07%**

总结

总的来说，该团队是第一个利用最先进的 3D 人体姿势评估技术，提供能够从步行视频中实时识别出情感状态的方法。值得一提的是，这个研究最终促成了一个视频数据集 —— EWalk，内容都是些人们的行走视频，被分别打上了对应的情感标签。

目前该方法当然也不是尽善尽美的，比如：

算法主要还是取决于 3D 人体姿势评估技术和步态提取算法的精度，换而言之，如果姿势或步态存在噪声，那么相应的情绪预测就可能是不准确的。

该情感算法需要提取全身关节的位置，一旦视频存在被遮挡的情况，就有可能无法获得全身的姿势数据。

行走动作必须是自然的，且不涉及任何配件（手提箱、手机……）

无论如何，这昭示着在机器读懂人类情绪这条道路上，已经取得了关键一步。在未来的世界里，机器人无需与我们产生交流，也能看透我们内心的小九九。所以，颤抖吧，人类！

- END -