TensorFlow平台上基于LSTM神经网络的人体动作分类

杨 煜，张 炜

（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要：随着人体运动数据采集技术的发展，基于数据的人体运动的研究越来越受到人们的关注。人体运动的研究在医疗康复、运动训练、虚拟现实、以及影视和游戏等领域有着很大的应用空间。人体动作分类就是基于大量已标注动作名称的人体动作，对未标注的人体动作进行分类标注。在本文中，研究提出了一种基于长短时记忆网络（LSTM）的人体动作分类模型。首先，将人体动作表示为时间序列的形式。然后，将人体动作序列逐帧输入到去掉输出层的正向和反向LSTM中，并将隐藏层输出依次送入Mean pooling层和逻辑回归层得到最终的分类结果。最后，研究利用目前流行的深度学习平台TensorFlow实现本次研发的分类模型并进行训练。基于此，又进一步利用人体动捕数据库HDM05的数据进行实验来验证提出的分类模型，经过训练，该模型在测试集上的分类准确率达到了94.84%。

关键词：人体动作分类；长短时记忆网络；时间序列；TensorFlow；HDM05

中图分类号：TP183 文献标志码：A 文章编号：2095-2163(2017)05-0041-05

Human action classification based on LSTM neural network on TensorFlow

YANG Yu, ZHANG Wei

(*School of Computer Science and Technology, Harbin Institute of Technology, Harbin* 150001, *China*)

Abstract: With the development of human motion data acquisition technology, the research of human motion based on data has attracted more and more attentions. The research of human motion has great application space in medical rehabilitation, sports training, virtual reality, film and television, games and so on. Human action classification aims to classify unlabeled human actions based on a large number of labeled human actions. This paper proposes a human action classification model based on Long Short-Term Memory network (LSTM). Firstly, represent human actions as a form of time series; then, input one human action by frame order into two LSTMs without output layer, one is forward LSTM and the other is backward LSTM, and pass the hidden layer outputs of LSTMs into the Mean pooling layer and the logical regression layer to get the final classification results; finally, implement the classification model and train it with the popular deep learning platform of TensorFlow. The research uses the data of human motion capture database HDM05 to validate the proposed classification model, and the accuracy rate of the classification model reaches 94.84% on test set.

Keywords: classification of human actions; LSTM; time series; TensorFlow; HDM05

1 概 述

随着人体运动数据采集技术的发展，基于数据的人体运动的研究越来越受到人们的关注。人体运动的研究在医疗康复、运动训练、虚拟现实、人机交互、以及影视和游戏等有着很大的应用空间。

人体运动可以表示为人体各部分在3D空间中的运动[1]，而人体动作可以看作是人体运动过程中的一个完整独立的动作片段，例如可以把屈膝、跳起、落地的这一段人体运动看作一个“跳跃”动作。人体动作的表示通常是基于各关节点的位置的[2]或基于身体各部分的旋转姿态的[3-4]。在本文中，研究将利用人体各部分的旋转姿态来表示人体动作，人体动作可以看作以一个时间序列[3,5-6]，序列中每一帧为身体各部分用四元数表示的旋转姿态。

人体动作分类问题是人体运动研究的重要问题之一。人体动作分类是基于大量已标注动作名称的人体动作，对未标注的人体动作进行分类标注。人们为解决人体动作分类问题应用了许多分类算法。随着神经网络的发展，许多研究者尝试用已经构建的神经网络模型进行人体动作的分类并取得了很好的效果。譬如Du等[7]利用分层级联的多个循环神经网络对人体动作进行分类。Cho和Chen[7]将人体动作序列的每一帧数据单独拿出来训练神经网络并进行分类，然后用投票法由各帧的分类结果得出序列分类的结果。Huang等[4]在将人体运动数据表示为李群的基础上，应用深度神经网络分类人体运动。

在本文中，研究构建了由双向LSTM神经网络和逻辑回归层组成的人体动作分类模型，并用TensorFlow平台实现模型的搭建和训练过程。TensorFlow是谷歌开源的数值计算平台，其中集成了大量神经网络模型的代码实现，使其成为了一个强大的深度学习平台。文献[8]中就是用TensorFlow实现的基于BP神经网络的手写字符识别方法。

作者简介：杨 煜（1992-），男，硕士研究生，主要研究方向：数据挖掘；张 炜（1975-），男，博士，副教授，主要研究方向：数据挖掘、无线传感器、数据分析等。

收稿日期：2017-06-06

在接下来的部分，先介绍人体动作分类的神经网络模型，再探讨论述了其在 TensorFlow 平台下的实现和训练，最后研究利用人体动作捕获数据库 HDM05[3] 的数据进行实验以验证模型的分类效果。

2 分类模型

2.1 人体动作分类问题概述

本文中讨论的人体动作分类问题是基于分割好的人体动作进行的，每个人体动作有唯一准确的动作类别标签。如前所述，人体动作分类就是基于大量已标注类别的动作，对未标注的动作进行分类标注。人体动作分类模型的训练和分类的过程如图 1 所示。

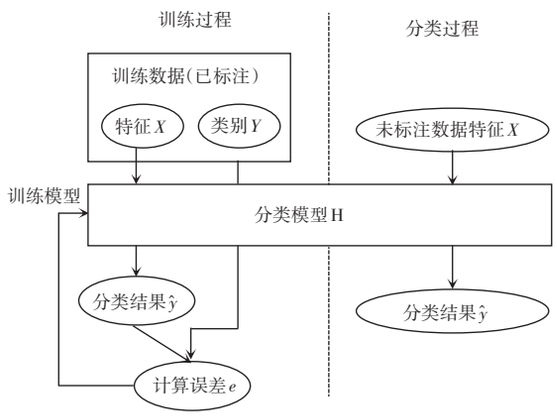

图1 人体动作分类模型的训练和分类过程

Fig. 1 The training and classification process of a human action classification model

对于分类问题，一般来说需要关注 2 个问题，即每条实例的数据形式，以及分类所用的算法或模型。在这里，首先介绍人体动作的数据表示，在后面的章节中重点深度剖析本文提出的分类模型及其在 TensorFlow 平台下的实现。

人体的结构和形态十分复杂，不同人的体态差异也很大，因此则需要用人体骨骼模型来对人体进行抽象。人体骨骼模型由抽象的骨头和关节构成，人体动作可以看做是人体骨骼模型中所有骨头的旋转姿态构成的一个时间序列。图 2 所示的是一个简单的包含 17 块骨头的人体骨骼模型，使用人体骨骼模型表示人体动作使得对人体运动的研究可以方便地迁移到不同人或骨骼模型上去。

人体骨骼模型并不是人体分类问题研究的一个限制因素。对于具体的研究问题和人体动作数据集，可以使用不同的人体骨骼模型进行表示，比如有的数据采集包含了手指上的运动，就要使用细化到手指的人体骨骼模型来替代用一个骨头表示手部运动的模型。

这里用时间序列来表示人体动作，假设有一个人体动作 X，则

$$X = (X^1, X^2, \cdots, X^t, \cdots, X^T) \tag{1}$$

其中，T 为人体动作的帧数即序列长度，每一帧表示在那一时刻人体骨骼模型上各骨头的旋转姿态，此处的旋转姿态用四元数表示，即对于 X 中的一帧数据 X^t，可以表示为：

$$X^t = (q_1^t, q_2^t, \cdots, q_m^t, \cdots, q_M^t) \tag{2}$$

其中，M 表示所用骨骼模型中骨头的块数，每一帧的数据由 M 个四元数组成。

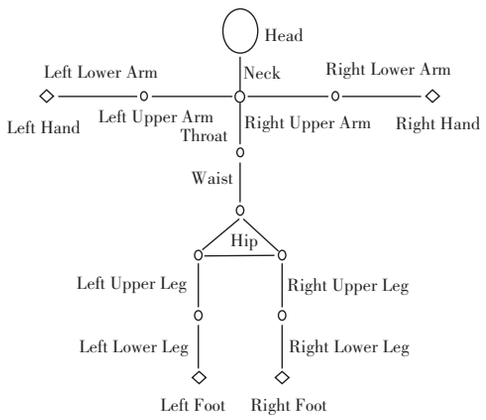

图2 人体骨骼模型

Fig. 2 Human skeleton model

基于上述人体动作的表示方法，每个人体动作被表示成一个多元时间序列。人体动作分类问题就可以看作是对这些时间序列进行分类的一个问题，同时用已标注的人体动作来训练本文的模型，并用训练好的模型对未标注的人体动作进行分类。

2.2 RNN 和 LSTM

讨论了人体动作的表示方法，接下来要讨论的就是图 1 中的分类模型部分。在本文中所用的长短时记忆网络(LSTM)是一种特殊的循环神经网络(RNN)。图 3 所示的是一个包含 2 个输入节点、3 个隐藏节点和 1 个输出节点的 RNN 网络。

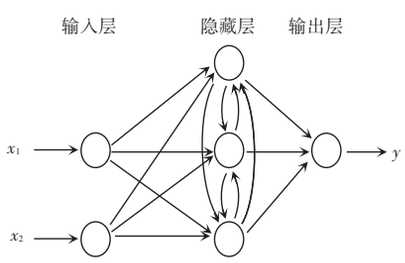

图3 一个简单的 RNN 网络示意图

Fig. 3 The sketch of a simple RNN network

RNN 与前馈神经网络(BPNN)的不同之处在于其中增加了同层隐藏层节点间的连线，因此被称为循环神经网络。RNN 通常用于处理序列问题，序列的每一帧依次从输入层传入网络，输入层节点个数等于序列每一帧的数据维数。隐藏层节点的计算不只依赖于当前输入层的输入，也依赖于上一时刻该隐藏层各节点的激活值。对于输入序列$(x^1, x^2, \cdots, x^T)$，RNN 网络将得到隐藏层序列$(h^1, h^2, \cdots, h^T)$和输出序列$(y^1, y^2, \cdots, y^T)$，具体的计算方法如下[2,10-11]：

$$h^t = H(W_h \cdot [x^t, h^{t-1}] + b_h) \tag{3}$$

$$y^t = O(W_{ho} \cdot h^t + b_o) \tag{4}$$

其中，H 和 O 表示隐藏层和输出层所用的激活函数，W_h 表示输入层和上一时刻隐藏层到当前隐藏层的权重矩阵，W_{ho} 表示隐藏层到输出层的权重矩阵，而 b_h 和 b_o 分别表示隐藏层和输出层的偏斜向量。

RNN 网络存在的问题是难以发现序列中的时间间隔较

长的帧之间的关系，因为任一帧的输入对后续隐藏层节点和输出层节点的影响会随着时间越来越小。LSTM 网络可以有效地解决这一问题，LSTM 在隐藏层加入了状态，使得网络对序列中较长时间前的输入有了记忆的能力。图 4 所示的是 LSTM 的隐藏层的结构示意图，当序列的第 t 帧输入到网络中时，LSTM 隐藏层的输入包括网络的当前输入 x^t，上一时刻的隐藏层输出向量 h^{t-1}，以及隐藏层状态 C^{t-1}。隐藏层的任务是计算并输出向量 h^t，并更新状态得到 C^t，为此隐藏层加入了遗忘门 f、输入门 i 以及输出门 o。遗忘门 f 决定状态 C 中的哪些信息被丢弃，输入门 i 决定输入由 x^t 和 h^{t-1} 得到的更新信息中有哪些能够用于状态 C 的更新，经过遗忘门和输出门，状态 C 的更新完成。然而 LSTM 中加入隐藏层状态的目的是使其对隐藏层的输出 h 产生影响，因此输出门 o 用于决定状态 C 中的如何作用到 h^t 的计算中。

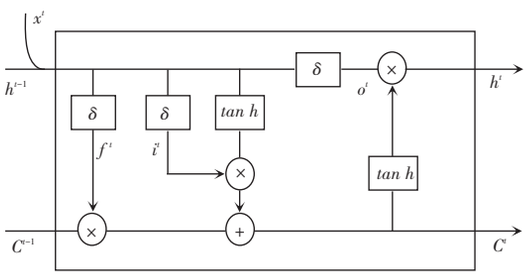

图 4　LSTM 隐藏层示意图

Fig. 4　The sketch of the hidden layer of LSTM network

图 4 中的 δ 表示的 sigmod 激活函数，3 个新加门以及隐藏层输出 h^t 和状态更新 C^t 的计算表达式如下：

$$f^t = \delta(W_f \cdot [x^t, h^{t-1}] + b_f) \tag{5}$$

$$i^t = \delta(W_i \cdot [x^t, h^{t-1}] + b_i) \tag{6}$$

$$o^t = \delta(W_o \cdot [x^t, h^{t-1}] + b_o) \tag{7}$$

$$C^t = tan\ h(W_c \cdot [x^t, h^{t-1}] + b_c) + f^t \cdot C^{t-1} \tag{8}$$

$$h^t = o^t \cdot tan\ h(C^t) \tag{9}$$

可以看出这 3 个门的输入都是 x^t 和 h^{t-1}，同时每个门中有自己的权重和偏斜。这些参数随着训练过程不断调优，在状态更新和隐藏层输出值的计算上发挥作用。

2.3　基于 LSTM 的人体动作分类模型

人体动作分类模型以人体动作即多元时间序列为输入，其输出为模型对该人体动作类别标签的估计。而 RNN 模型和 LSTM 模型的输出是一个与输入序列等长的时间序列，这显然和论文研究的问题是不一致的。为此，需要在 LSTM 模型的基础上进行改动使之适应人体动作分类问题。

研究中选用的基于 LSTM 的人体动作分类模型如图 5 所示。图 5 中，将分类模型按时间展开，图中标有 LSTM 的矩形表示的是 LSTM 隐藏层，其中同一行的 LSTM 矩形表示的是同一个 LSTM 隐藏层，这里只是将其按时间展开。

过程中，将 LSTM 网络原本的输出层去掉，并将隐藏层的输出 $(h^1, h^2, \cdots, h^T)$ 输入到一个平均池层，得到不再包含时间信息的向量 h，即：

$$h = \frac{1}{T}\sum_{t=1}^{T} h^t \tag{10}$$

至此，将去掉输出层并加入平均池层的 LSTM 网络称为一个 mLSTM，同时，还要用到一个 mLSTM 的反向版本 Bi-mLSTM，Bi-mLSTM 与 mLSTM 的结构相似，只是 Bi-mLSTM 要求输入序列按照时间的反向顺序输入，Bi-mLSTM 的输出用 h_b 表示。最后，设计组合了一个逻辑回归模型，逻辑回归层以 mLSTM 的输出 h 和 Bi-mLSTM 的输出 h_b 连接成的新向量为输入，并通过 softmax 层输出分类的结果。

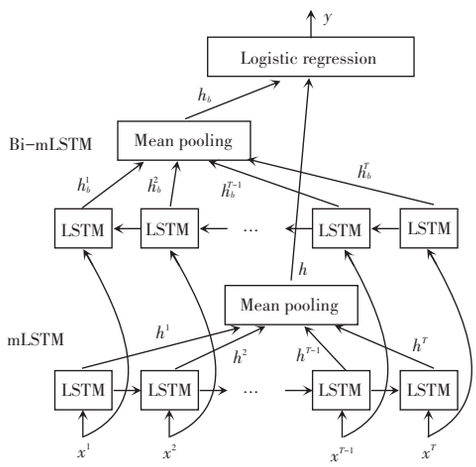

图 5　基于 LSTM 的人体动作分类模型

Fig. 5　Human action classification model based on LSTM network

3　在 TensorFlow 平台上的模型实现

3.1　TensorFlow 平台介绍

TensorFlow 是谷歌推出的第二代人工智能学习系统，而且有着很多优秀的特点，对其阐释如下：

1) 高度的灵活性。TensorFlow 不仅能够用于搭建并训练各种神经网络模型，还可以完成很多其他计算任务，用户只需要将自己的计算模型设计成数据流图的形式就可以应用 TensorFlow 完成任务。

2) 可移植性强。TensorFlow 可以在 CPU 和 GPU 上运行，这即使其能够移植到台式机、服务器和手机等许多设备上。

3) 提供了大量机器学习的模型，使得科研和开发人员可以省去重写底层实现的繁琐工作。

4) 自动求微分。对于使用梯度下降法进行训练的机器学习模型，用户只需要定义损失函数以及模型中哪些参数是可训练的，TensorFlow 就能够自动求微分导数并用梯度下降法训练模型参数。

5) 性能优化。对于多 CPU 和 GPU 的工作平台，TensorFlow 能够很好地支持多线程、队列、异步操作等。

在 TensorFlow 下，可以使用 python 或 C++的代码来搭建数据流图进行计算。流图中的节点表示数学操作，线表示在节点间传递的数据张量即多维数据数组。

用 TensorFlow 实现模型一般分为构建数据流图、训练模型、使用模型这 3 个阶段。在 TensorFlow 中，可以用常量、变量、以及操作来构建数据流图。其中，变量包括输入变量、可训练的变量以及其他变量。在流图中加入输入变量需要用占位符 placeholder 占位，之后在训练和使用模型时用 feed 操作将数据从 placeholder 输入到模型中。可训练的变量用来表示模型中的权重和偏移等参数，在构建这些变量时需要设置

trainable = True。在训练阶段，可以调用训练相关的操作使这些模型参数随着训练数据得到训练。

3.2 构建 TensorFlow 流图实现人体动作分类模型

在 TensorFlow 中提供了 LSTMCell 操作来支持 LSTM 模型的搭建。LSTMCell 相当于 LSTM 模型的隐藏层，在内部封装了 LSTM 隐藏层包含的遗忘门、输入门和输出门等结构，同时还可根据研究需要设置隐藏层结点个数。

在用 TensorFlow 搭建神经网络的过程中，不再以神经网络中的节点为单位进行布局，而是以层为基础来考虑。因为像 LSTMCell 这样的 TensorFlow 操作直接代表了网络中的一个隐藏层。因此包含多个节点的输入层和输出层也都用向量的形式来表示，向量长度即为该层节点的个数。

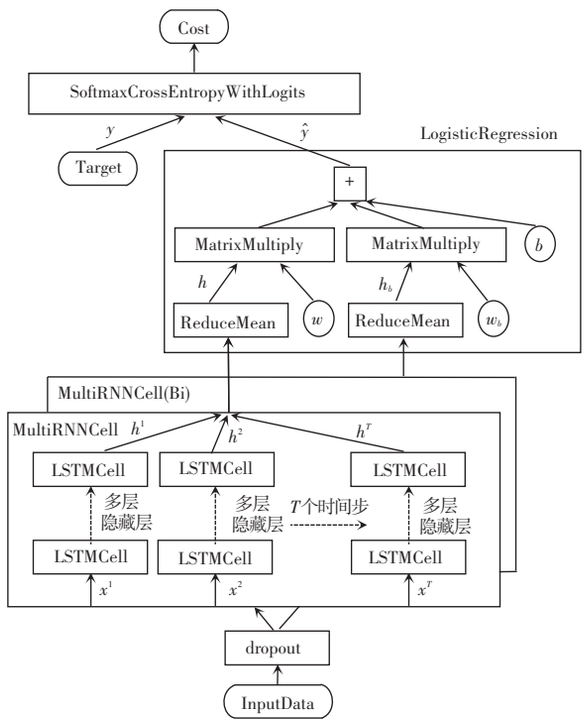

图 6　人体动作分类模型的 TensorFlow 流图

Fig. 6　TensorFlow flow diagram of human action classification model

研究构建的 TensorFlow 流图如图 6 所示。图中胶囊形单元表示模型的输入和输出，矩形单元表示 TensorFlow 中的操作，圆形单元表示可训练的模型参数。在该数据流图中，InputData 是一个人体动作实例，即一个多元时间序列。输入数据 InputData 经过 dropout 操作，dropout 操作的目的是防止模型过于拟合。接下来，数据被传入 2 个 MultiRNNCell 中，MultiRNNCell 是 TensorFlow 提供的 RNN 的主要操作，相当于 RNN 的整个隐藏层。图中的 MultiRNNCell 中的内容是其按时间展开图，隐藏层用 LSTMCell 实现，其中可以包含多个隐藏层。在图中省略了反向 MultiRNNCell 的详细内容，因为 2 个 MultiRNNCell 的结构相同，只是在输入序列数据时一个按照正常顺序输入，另一个按照相反的顺序进行输入。2 个 MultiRNNCell 得到的输出序列分别经过 ReduceMean 操作得到与时间无关的平均向量 h 和 h_b，向量的长度即为隐藏层节点个数。最后，h 和 h_b 经过一个手动构建的逻辑回归层和 *softmax* 激活函数，并用交叉熵损失函数来计算模型输出与真实的类别标签的误差。

以上就是本次研究利用 TensorFlow 搭建的基于 LSTM 的人体动作分类模型。模型中的主要训练参数包括输入层到隐藏层的权重和偏斜、LSTMCell 中 3 个门的权重和偏斜、以及逻辑回归层的权重和偏斜。给出损失函数 *Cost* 后，使用 TensorFlow 提供的训练操作可以自动求 *Cost* 关于每个参数的微分导数并用梯度下降法对模型进行训练。

4 实验

4.1 实验数据

综上研究论述后，即将用 HDM05 动作捕获数据库[3] 中的数据进行实验以检验分类模型的效果。HDM05 中有 2 337 条切分好的人体动作数据，每个人体动作都标注了类别标签，共有 130 个类别。HDM05 的人体动作数据采集了人体 31 个部分的运动数据，其网站上提供了将这些数据转换成旋转姿态四元数的代码。

经过观察，进一步发现 HDM05 数据库中有些骨头上的姿态四元数固定不变，比如左右肩的四元数，为此选择抛弃这些部分的数据不用，以免影响模型的效果。另外，由于这 130 个动作类别都和头颈的运动无关，因此头和脖子的数据也可以舍弃不用。最终，就实际确定了包括 15 个骨头的数据进行模型的训练和动作的分类，具体来说则分别是：左大腿、左小腿、左脚、右大腿、右小腿、右脚、腰下部、腰上部、胸、左大臂、左小臂、左手、右大臂、右小臂、右手。

对于 HDM05 中的 130 个类别标签有很多类别应属于相同的动作，比如 jogging starting from air 和 jogging starting from floor，jogging 2 steps 和 jogging 4 steps[3]。文献 [7] 中将这 130 个类别合并成 65 个类别，在此基础上文献 [2] 指出有些类别仍难以区分，比如 deposit 和 grab，这 2 个类别需要细化到手指的动作才有可能区分，sitDownChair 和 sitDownTable 在只有人体运动数据的情况下也难以识别桌子和椅子的不同。最终，本次研究就将文献 [7] 中给出的 65 个类别合并成了 54 个类别进行人体动作的分类实验，例如 kickLFront 和 kickLSide 合并，jogOnPlace 和 run 合并，deposit 和 grab 合并等。

由于每个人体动作的时间长度不一，最长的动作长度为 901 帧，还要将每个人体动作放缩到统一长度为 256 帧。对于不足 256 帧的人体动作，就需要在动作的末尾用全零的帧将其补齐到 256 帧；对于长度超过 256 帧的人体动作，将会在其中随机不重复地选取 256 帧，并使其按照原来的顺序构成缩短后的序列。

4.2 参数设置

输入数据的每一帧包含 15 块骨头上的旋转四元数，因此模型的输入节点可设置为 60 个。输出层节点设置为 54 个，与所有 54 个动作类别相对应。序列长度设为 256，与研究规定的人体动作统一长度一致。其他参数的设置将在表 1 中给出清晰呈现。

模型中的训练参数的初始化会对训练效果产生很大的影响，这里就选用 TensorFlow 提供的 *random_uniform_initializer* 对逻辑回归层的训练参数进行初始化，并用 *orthogonal_initializer* 方法对 LSTMCell 中的遗忘门、输入门和

输出门的参数进行初始化。此外，实践证明在新建 LSTMCell 时将参数 forget_bias 从默认的 0 调整为 1.0 会使模型的训练效果产生有所提升。研究将使用批量随机梯度下降法进行训练，也就是每次将 4 条训练实例一同输入给模型对模型进行训练。

表 1　参数设置

Tab. 1　Parameter setting

参数名称	参数值
输入层节点个数	60
输出层节点个数	54
序列长度	256
隐藏层层数	1
每层隐藏层节点个数	128
学习率	0.1
每次训练的实例数	4
forget_bias	1.0

4.3　实验结果

实验过程中，将 HDM05 中的 2 337 个人体动作按类别标签排序，然后在每相邻的 15 个动作中随机选取 1 个人体动作放入测试集，并将其他人体动作放入训练集。这样做保证了训练集和测试集的类别分布一致。而后，用训练集的全部动作迭代训练模型 50 次，每次迭代会将训练集数据随机重排列。为此，则记录了每次迭代后模型的损失函数值以及模型在训练集和测试集上的分类准确率，记录结果如图 7 和图 8 所示。从图中可以看出损失函数值随着迭代而下降，而分类的准确率随着迭代而上升，最终两者的变化都将趋于平稳，这也符合神经网络模型的一般训练过程。在 50 次迭代的过程中，模型在训练数据上的准确率最高达到 98.44%，在测试数据上的准确率最高达到 94.84%。

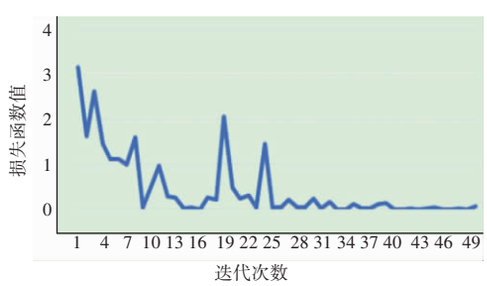

图 7　模型的损失函数值随迭代次数的变化曲线

Fig. 7　The curve of the loss function with the number of iterations

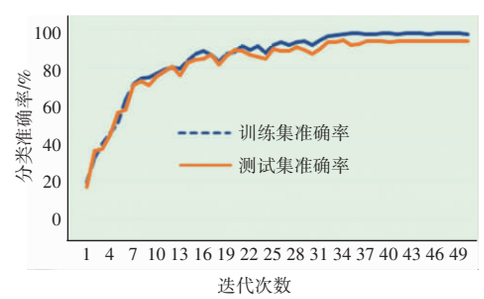

图 8　分类准确率随迭代次数的变化曲线

Fig. 8　The curve of the accuracy with the number of iterations

5　结束语

在本文中，研究提出了一种基于 LSTM 神经网络的人体动作分类模型。通过将人体动作表示为时间序列的形式，序列上的每一帧由人体各部分的旋转姿态四元数构成。接着将人体动作序列逐帧输入到去掉输入层的正向和反向 LSTM 中，并将隐藏层输出送入 Mean pooling 层关于时间求平均，再将 Mean pooling 层的输出送入逻辑回归层得到最终的分类结果。

之后，又使用 TensorFlow 搭建了设计研发的分类模型，利用 TensorFlow 平台提供的 LSTMCell 等操作将模型构建成数据流图的形式，并用 TensforFlow 自动计算微分导数的功能选取梯度下降法训练模型。研究最后，则利用 HDM05 人体动作捕获数据库的数据进行实验验证了模型的分类效果，就是将 HDM05 的数据随机划分为训练集和测试集，用训练集训练模型后，该模型在测试集上的分类准确率达到了 94.84%.

参考文献:

[1] YE M, ZHANG Q, WANG L, et al. A survey on human motion analysis from depth data [M] // GRZEGORZEK M, THEOBALT C, KOCH R, et al. Time-of-flight and depth imaging. sensors, algorithms and applications. Lecture Notes in Computer Science. Berlin: Springer ,2013: 149-187.

[2] DU Yong, WANG Wei, WANG Liang. Hierarchical recurrent neural network for skeleton based action recognition [C] //2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) . Boston, MA, USA: IEEE,2015: 1110-1118.

[3] SEMPENA S, MAULIDEVI N U, ARYAN P R. Human action recognition using dynamic time warping [C] //International Conference on Electrical Engineering and Informatics, Iceei 2011. Bandung, Indonesia: IEEE, 2011: 1-5.

[4] HUANG Zhiwu, WAN Chengde, PROBST T, et al. Deep learning on lie groups for skeleton-based action recognition [J] . arXiv preprint arXiv: 1612.05877,2016.

[5] GONG Dian, MEDIONI G, ZHAO Xuemei. Structured time series analysis for human action segmentation and recognition [M] . IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014,36 (7) : 1414-1427.

[6] LI Kang, FU Yun. Prediction of human activity by discovering temporal sequence patterns [J] . IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 36(8) : 1644-1657.

[7] CHO K, CHEN X. Classifying and visualizing motion capture sequences using deep neural networks [C] // International Conference on Computer Vision Theory and Applications. Lisbon, Portugal: IEEE, 2014: 122-130.

[8] 张俊，李鑫. TensorFlow 平台下的手写字符识别 [J] . 电脑知识与技术，2016，12(16) : 199-201.

[9] MÜLLER M, RÖDER T, CLAUSEN M, et al. Documentation mocap database HDM05 [R] . Bonn: Universität Bonn, 2007.

[10] GRAVES A. Supervised sequence labelling with recurrent neural networks [M] . Berlin: Springer, 2012.

[11] GRAVES A, MOHAMED A, HINTON G. Speech recognition with deep recurrent neural networks [C] //Acoustics, Speech and Signal Processing (ICASSP) , 2013 IEEE International Conference on. Vancouver, BC, Canada: IEEE, 2013, 38(2003) : 6645-6649.