Computer Science and Application 计算机科学与应用, 2020, 10(5), 927-934
Published Online May 2020 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2020.105095

Hans汉斯

Action Prediction Research Based on Knowledge Distillation

Xiang Wang

Ocean University of China, Qingdao Shandong
Email: xiangwnp@foxmail.com

Received: Apr. 23rd, 2020; accepted: May 7th, 2020; published: May 14th, 2020

Abstract

Action recognition is a hot topic in the domain of computer vision, and it's widely applied in human-computer interaction, studio entertainment, automatic drive, intelligent video surveillance, and intelligent medical care. Action prediction is a special class of action recognition. Different from conventional action recognition which aims at recognizing complete actions, the purpose of action prediction is to distinguish an action before it's fully executed so that some objectives, such as accident early warning and crime prevention, can be achieved by analyzing the possible impact of the action. In order to solve the problem of real-time action prediction, this paper develops a multi-stage LSTM architecture that leverages knowledge distillation technique. The context-aware feature and action-aware feature are exploited for action modeling. The proposed multi-stage LSTM architecture is composed of two stages. In the first stage it focuses on the global, context-aware information. The second stage then combines these context-aware features with action-aware ones. In order to improve the performance of proposed method in the early stage, the knowledge distillation technique is exploited for transferring the knowledge from teacher model to student model. A novel loss function is designed for the whole action prediction architecture and the performance is improved with the novel loss function. Experimental results on the UT-Interaction dataset, JHMDB-21 dataset and the UCF-101 dataset show that the proposed methods not only improve the accuracy of action prediction but also have the ability of real-time running.

Keywords

Action Recognition, Action Prediction, Fall Prediction, Knowledge Distillation

基于知识蒸馏的实时动作预测方法研究

王　祥

中国海洋大学，山东　青岛
Email: xiangwnp@foxmail.com

文章引用: 王祥. 基于知识蒸馏的实时动作预测方法研究[J]. 计算机科学与应用, 2020, 10(5): 927-934.
DOI: 10.12677/csa.2020.105095

收稿日期：2020年4月23日；录用日期：2020年5月7日；发布日期：2020年5月14日

摘　要

动作预测是一类特殊的动作识别问题，不同于针对完整动作的传统动作识别，动作预测旨在动作尚未完成时尽可能早地识别动作所属的类别，以便对该动作可能造成的影响进行分析，从而实现事故预警、智能陪护、犯罪预警等目标。本文针对实时动作预测问题提出一种应用知识蒸馏技术的多阶段LSTM实时动作预测方法。本文中的动作预测模型为两阶段的LSTM模型，在第一阶段利用全局特征对动作进行分析，第二阶段利用全局特征与动作特征对动作进行分析。为提高动作预测模型的性能，本文利用知识蒸馏技术并设计新型的损失函数提高动作预测模型的性能。UT-Interaction数据集、JHMDB-21数据集以及UCF-101数据集的实验结果表明本文所提出的动作预测方法不但具有良好的动作预测能力，而且能够满足实际应用中的实时性要求。

关键词

动作识别，动作预测，跌倒预测，知识蒸馏

1. 引言

随着人工智能技术的快速发展，作为计算机视觉领域热门研究方向的动作预测技术在自动驾驶、智能视频监控、人机交互、智能医疗看护等多个领域具有十分广阔的应用前景。

动作预测是指实时对输入的视频序列进行分析处理，从而在该视频中所包含的动作尚未执行完之前尽可能早地对其动作类别进行识别。因此动作预测是对正在进行中的动作进行识别，属于一类特殊的动作识别技术。传统动作识别技术是对视频序列中已经完成的动作进行识别，而动作预测技术则是对视频序列中正在发生的动作进行识别。因此，动作预测技术与传统动作识别技术的不同之处在于视频序列中动作的完整性。在实际场景中，某些动作在动作早期在外观上具有相似性，例如“拥抱”和“握手”这两个动作在动作开始时都存在手臂前伸的举动，动作外观的相似性导致从部分视频序列中分析提取的特征是相似的，两个动作特征向量之间的距离较小，使预测模型无法有效对上述两个动作进行识别，增加了预测问题的难度。在对视频序列的观测结束之前，预测算法无法获取动作执行完毕所需要的时间，不能够确定动作的完成程度，无法通过动作持续时间的不同对动作进行识别。因此，从已经观测的部分视频序列中所提取的动作特征往往既不能提供用于识别这些动作的关键信息，也不能直接用于获取完整动作的时序结构。因此，与动作识别技术相比，动作预测技术具有关键动作特征信息缺乏和完整时序动作结构未知这两个特点，因此动作预测技术与动作识别技术相比更加具有挑战性。

为解决动作预测中动作早期可利用关键信息较少的难点，本文设计了基于知识蒸馏的多阶段 LSTM 动作预测模型，第一阶段考虑全局特征，第二阶段综合考虑全局特征与动作特征，充分利用上述两种特征对动作进行预测。为增加可用的关键信息，利用知识蒸馏技术将知识迁移到动作预测模型中，从而增强动作预测模型的性能。为充分发挥所设计架构的性能，设计了合适的损失函数以提高模型的预测性能。

2. 相关技术

2.1. 动作预测技术

与非实时动作预测技术不同，实时动作预测技术需要在任意时刻给出动作所属的标签，因此实时动作预测任务比非实时动作预测任务难度更高。在实时动作预测方法中，LSTM 模型能够输出任意时刻的动作概率，因此通常作为实时动作预测领域中的基础模型。Ma 等人[1]利用 LSTM 模型对时序动作进行建模，该模型利用卷积神经网络对每一帧图像进行特征提取，所提取的特征经过一个全连接层被送入到 LSTM 网络中，并输出该时刻图像中内容属于各个动作类别的概率。该方法中为了使所提出的方法能够更好的胜任动作预测任务，设计了一种名为 Ranking 的损失函数，将判断为正确类别的概率随时间的增长而增长这一先验知识引入到损失函数中，利用所设计的损失函数，提高了 LSTM 模型的预测性能。Aliakbarian 等人[2]同样采用 LSTM 模型对动作进行建模，在他们的方法中利用两阶段的 LSTM 模型对动作进行建模分析，第一阶段使用环境特征对动作进行建模，第二阶段使用 CAM 动作特征以及第一阶段的输出对动作建模。在他们的方法中同样设计了一种新型损失函数，该损失函数能够提高 LSTM 模型在动作初期的识别能力。

2.2. 知识蒸馏

知识蒸馏(Knowledge Distillation)是将复杂模型(Teacher)中的知识迁移到简单模型(Student)中。其中，Teacher 模型具有强大的能力而 Student 模型则更为紧凑。通过知识蒸馏，希望 Student 尽可能逼近抑或是超过 Teacher，从而用更少的复杂度的模型获得更好的预测效果。知识蒸馏概念最早由 Hinton [3]提出，通过引入 Teacher 的软目标以诱导学生网络的训练。Romero 等[4]提出最小化 Teacher 模型与 Student 模型输出之间的均方误差来达到知识迁移的目的。Yim 等[5]借助于 Gram 矩阵损失函数实现知识蒸馏，提高模型图像识别的能力，教师模型由 32 层网络组成，学生模型由 12 层网络组组成，利用 Gram 矩阵实现知识的迁移，将知识从教师模型迁移到学生模型中。Li 等[6]人证明了最下化 Gram 矩阵损失函数等价于最小化 MMD 损失函数[7]。在本文中同样借鉴知识蒸馏的思想，将从完整性视频中学习到的知识迁移到动作预测模型中。

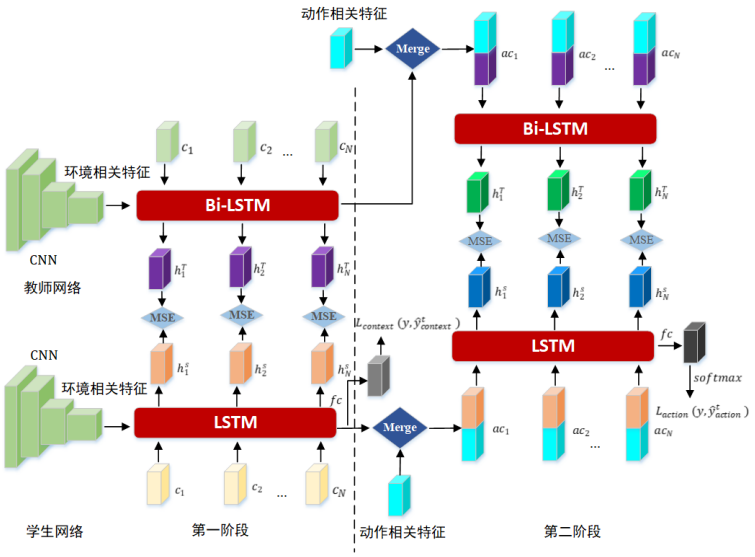

Figure 1. The architecture of action prediction network
图 1. 实时动作预测网络架构图

3. 基于知识蒸馏的动作预测模型

3.1. 网络架构

基于知识蒸馏的多阶段 LSTM 动作预测网络架构如图 1 所示，与现有的模型[8] [9] [10] [11]相比，本文方法无需将视频预先进行分割。该模型对动作的建模分为两阶段。第一阶段，关注动作所处的环境特征，将环境感知网络所提取的全局特征依次输入到模型中，计算关注环境因素的隐藏层特征。第二阶段，将第一阶段计算的环境相关的隐藏层特征与动作特征相融合，综合考虑全局特征与动作相关特征，利用上述两种特征推测动作的类别。

在每一阶段，网络都由一个教师(Teacher) - 学生(Student)学习模块组成，该学习模块用于将知识从教师网络迁移到学生网络，从而增强学生网络的性能。其中，学生网络由一层标准的长短期记忆单元(LSTM) [12]组成，对执行到不同程度的动作进行实时预测，尽可能早的预测动作所属的类别。教师网络由一层双向长短期记忆单元(Bi-LSTM) [13]组成。

3.2. 特征提取

在本文中，分别基于 RGB 图像数据以及人体骨骼点坐标数据提取全局特征以及动作特征，并且利用上述两种特征对动作进行建模。下面分别对两种特征的提取进行详细介绍。

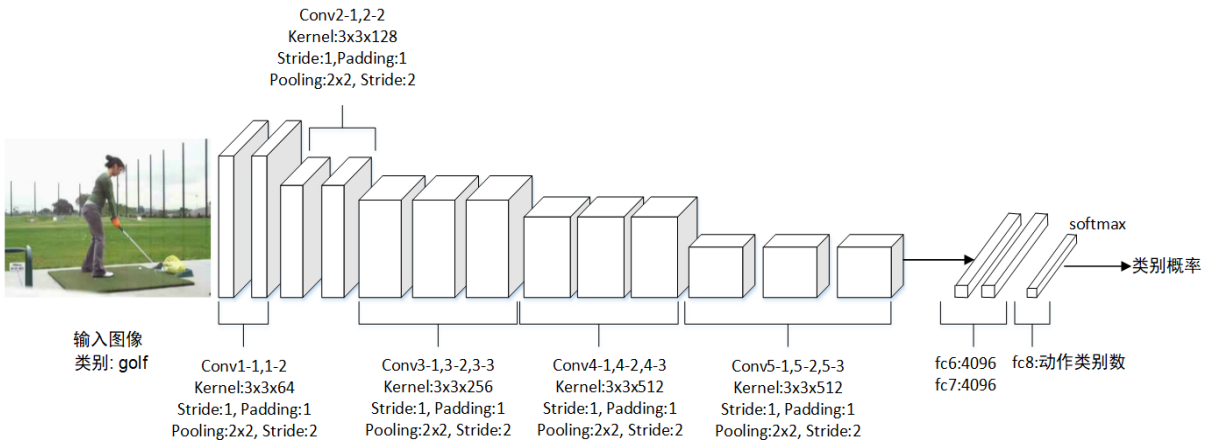

Figure 2. The global feature extraction network
图 2. 全局特征提取网络

3.2.1. 全局特征提取

如图 2 所示，本文中采用在 ImageNet 数据集上经过预训练的 VGG-16 网络[14]进行全局特征提取。该网络由五个卷积模块、五个池化层以及三个全连接层组成，各层的具体参数如图 2 所示。本文全局特征提取网络在 VGG-16 网络的基础上将 VGG-16 网络的 fc8 全连接层的神经元个数改为具体数据集中的动作类别数，以 fc7 全连接层的输出作为全局特征，由图 2 可知，全局特征的维度为 4096。

3.2.2. 动作特征提取

在本文中，采用 OpenPose [15]提取人体骨骼点坐标。基于骨骼数据的动作特征提取网络如图 3 所示。首先利用 OpenPose 对每一帧中的人物进行骨骼点坐标提取，然后将每一帧表示为一个三维张量，张量的维度为 $T \times N \times D$，其中第一个维度代表帧数(本文中 $T = 1$)，第二个维度代表骨骼数(本文中 $N = 18$)，第三个维度代表骨骼点的坐标维度(本文中 $D = 2$)。动作特征提取分为三个阶段，第一阶段为点特征提取，

第二阶段为全局特征提取，第三阶段为时序依赖特征提取。在第一阶段，输入张量首先经过两次 1 × 1 (conv1, conv2)卷积操作，会强迫网络学习单个骨骼点坐标之间的相互关系，进而得到点级别的特征表示。随后，将特征图按参数(0, 2, 1)进行转置操作，将各个骨骼点从第二个维度转换到第三个维度。在第二阶段，使网络学习到各个骨骼点之间的相互关系，得到动作的全局特征。在第三阶段，将第二阶段输出的特征图展成一个 256 维的向量，将该向量输入到 LSTM 单元中，得到动作在时间维度上的依赖关系。最后 LSTM 隐藏层的输出经过一个全连接层完成最终的动作识别任务。网络训练完成后，将 LSTM 单元隐藏层的输出作为动作特征，其中 LSTM 隐藏层的维度为 256。

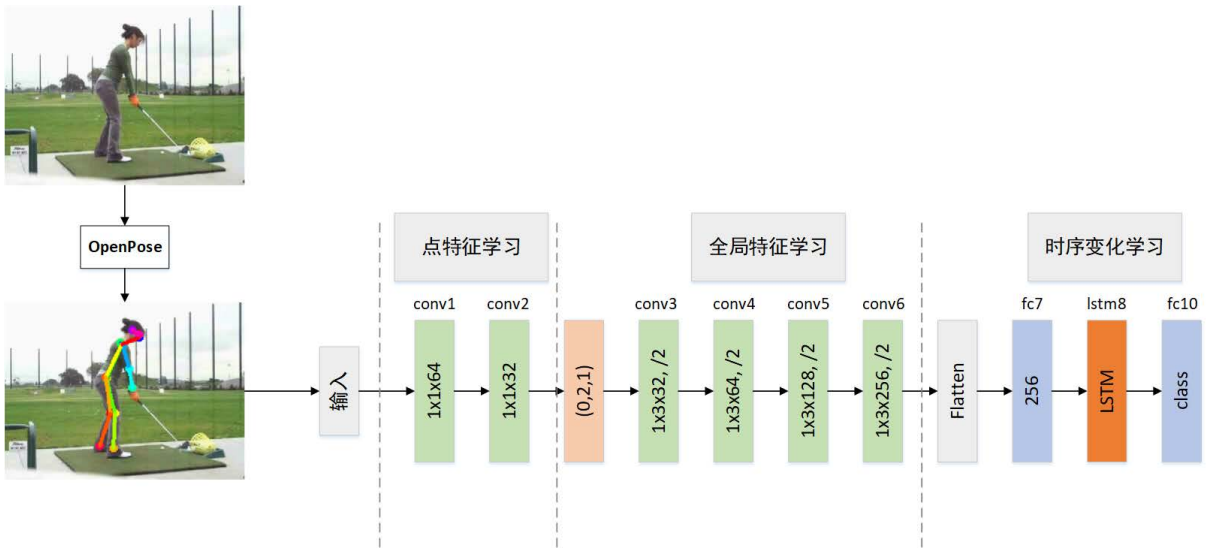

Figure 3. The action feature extraction network
图 3. 动作特征提取网络

3.3. 损失函数

在动作预测领域，一个精巧设计的损失函数可以提高动作预测模型的性能。借鉴[2] [11]中的损失函数设计思想，本文针对所提出的包含知识蒸馏的多阶段 LSTM 动作预测模型设计了一个高效的损失函数，如公式(1)所示。

$$L\left(y,\hat{y}\right)=L_C\left(y,\hat{y}\right)+L_{TS}\left(S,T\right) \tag{1}$$

本文中的动作预测模型分为两个阶段，模型每个阶段的损失函数都如公式(1)所示，但每个阶段损失函数的参数不同。公式(1)中，L_C 表示模型的预测损失，y 表示样本的标签，$\hat{y}$ 表示模型的预测值，L_{TS} 表示教师网络与学生网络的知识蒸馏损失，S 表示学生网络计算的各个时刻的特征，T 教师网络计算的各个时刻的特征。下面对 L_C 以及 L_{TS} 做具体的介绍。

$$L_C\left(y,\hat{y}\right)=-\frac{1}{N}\sum_{k=1}^{N}\sum_{t=1}^{T}\left[y^t\left(k\right)\log\left(\hat{y}^t\left(k\right)\right)+\frac{t\left(1-y^t\left(k\right)\right)}{T}\log\left(1-\hat{y}^t\left(k\right)\right)\right] \tag{2}$$

公式(2)中，$y^t\left(k\right)$ 表示在 t 时刻动作的标签，例如 $y^t\left(k\right)=1$ 表示样本属于类别 k，$y^t\left(k\right)=0$ 表示样本不属于类别 k。$\hat{y}^t\left(k\right)$ 表示模型对样本的预测标签，N 表示动作类别综述，T 表示输入序列的长度(帧数)。

$$L_{TS}\left(S,T\right)=\alpha L_{MSE}\left(S,T\right)=\alpha\left\|S-T\right\|_F^2 \tag{3}$$

公式(3)为知识蒸馏损失函数，该损失函数用于指导学生网络的学习，使学生网络输出的隐藏层特征尽可能逼近或等于教师网络输出的隐藏层特征。公式 3 中，S 为学生网络在所有时刻输出的隐藏层特征，T 为教师网络在所有时刻输出的隐藏层特征，α 表示该损失函数的影响因子。

4. 实验结果

在本节中，给出在 UT-Interaction 数据集、JHMDB-21 数据集以及 UCF-101 数据集上本文提出方法与现有相关方法的比较结果，采用折线图以及表格的方式对观测比率为 0.1、0.2、0.3、0.4、0.5、0.6、0.7、0.8、0.9、1.0 时各方法的识别准确率进行展示。

4.1. JHMDB-21 数据集

在 JHMDB-21 数据集上，本文方法与 DP-SVM [16]、S-SVM [16]、Where/What [17]、MS [2]等方法的比较结果如图 4 所示。由图 4 可知，本文所提出方法的识别准确率在所有观测比率上都远远优于所比较的方法。

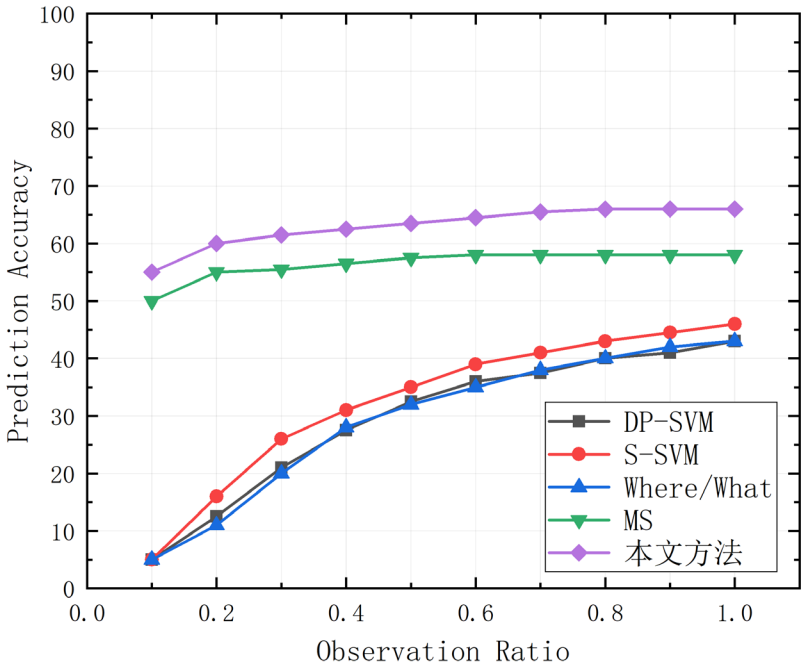

Figure 4. The result in JHMDB-21 dataset
图 4. JHMDB-21 数据集实验对比结果

4.2. UT-Interaction 数据集

在 UT-Interaction 数据集上，本文方法与 IBoW [16]、DBoW [16]、SC [18]、MSSC [18]、MMAPM [9]的比较结果如图 5 所示。由图 5 可知，本文方法在所有的观测比率上均超过所比较的方法。

4.3. UCF-101 数据集

UCF-101 数据集是十分具有挑战性的数据集，在 UCF-101 数据集上，本文方法与 IBoW [16]、DBoW [16]、DeepSCN [8]、Mem-LSTM [10]、MTSSVM [9]、MSSC [18]、MS [2]、PTSL [11]的比较结果如图 6 所示。由图 6 可知，除 PTSL 方法外，本文方法在所有观测比率上皆超过其他几种方法。

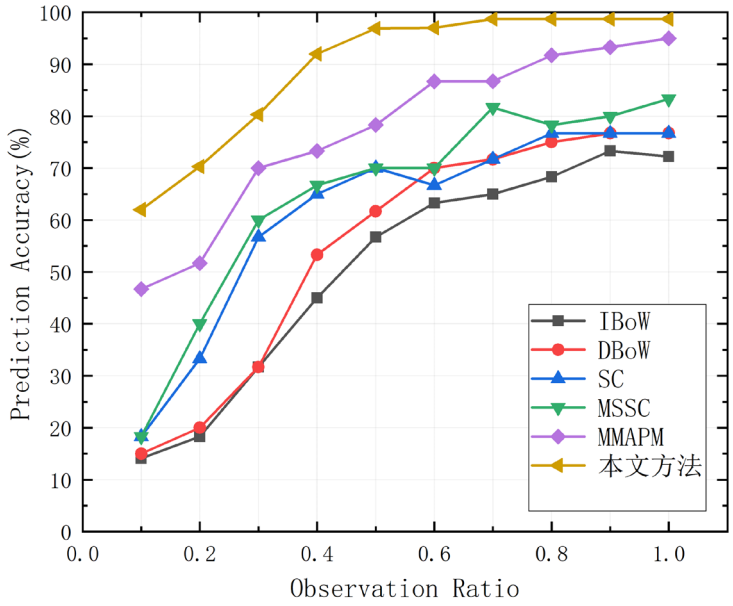

Figure 5. The result in UT-Interaction dataset
图 5. UT-Interaction 数据集实验对比结果

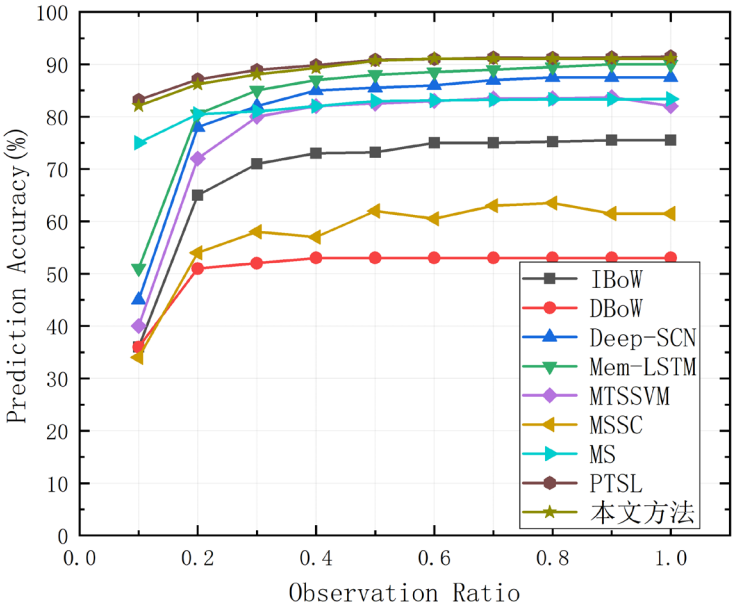

Figure 6. The result in UT-Interaction dataset
图 6. UT-Interaction 数据集实验对比结果

5. 总结与展望

本文针对实时动作预测问题提出一种应用知识蒸馏技术的多阶段 LSTM 实时动作预测方法。本文中的动作预测模型为两阶段的 LSTM 模型，在第一阶段利用全局特征对动作进行分析，第二阶段利用全局特征与动作特征对动作进行分析。为提高动作预测模型的性能，本文利用知识蒸馏技术将教师模型从完整视频序列中所学到的知识迁移到学生模型中，提高学生模型的性能。为更好地发挥本文所设计模型的性能，本文针对所设计的模型架构设计了适用于本文模型架构的损失函数，利用该损失函数可以使模型

得到更好的训练效果。实验结果表明本文所提出的动作预测方法不但具有良好的动作预测能力，而且能够满足实际应用中的实时性要求。

虽然本文在实时动作预测方面提出了较为不错的方法，但在对实时动作预测的研究工作中仍然存在一些不足之处有待进一步完善，本文中的动作特征借助于人体骨骼点数据，利用 OpenPose 从图像中提取人体骨骼点数据，但在某些情况下图像中的人物是被遮挡的以及图像中人物较小，因此在上述情况下无法提取人体动作特征，从而导致方法的准确性降低。因此可以在本文工作的基础上，从图像中提取动作特征，然后将两种动作特征融合组成新的动作特征，从而解决无法提取骨骼点坐标问题。

参考文献

[1] Ma, S., Sigal, L. and Sclaroff, S. (2016) Learning Activity Progression in LSTMs for Activity Detection and Early Detection. *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 1942-1950. https://doi.org/10.1109/CVPR.2016.214

[2] Aliakbarian, M.S., Saleh, F.S., Salzmann, M., Fernando, B., Petersson, L. and Andersson, L. (2017) Encouraging LSTMs to Anticipate Actions Very Early. Proc. *IEEE International Conference on Computer Vision*, Venice, 22-29 October 2017, 280-289. https://doi.org/10.1109/ICCV.2017.39

[3] Hinton, G., Vinyals, O. and Jeff, D. (2015) Distilling the Knowledge in a Neural Network. *NIPS* 2014 *Deep Learning Workshop*, Montreal, 8-13 December 2014, 1546-1552.

[4] Adriana, R., Gatta, C. and Bengio, Y. (2015) FitNets: Hints for Thin Deep Nets. 3*rd International Conference on Learning Representations*, San Diego, 7-9 May 2015, 1-13.

[5] Yim, J. (2017) A Gift from Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning. *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 4133-4141.

[6] Li, Y. and Wang, N. (2017) Demystifying Neural Style Transfer. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, Melbourne, 19-25 August 2017, 2230-2236.

[7] Gretton, A., *et al*. (2012) A Kernel Two-Sample Test. *Journal of Machine Learning Research*, **13**, 723-773.

[8] Kong, Y., Tao, Z. and Fu, Y. (2017) Deep Sequential Context Networks for Action Prediction. *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 3662-3670. https://doi.org/10.1109/CVPR.2017.390

[9] Kong, Y., Kit, D. and Fu, Y. (2014) A Discriminative Model with Multiple Temporal Scales for Action Prediction. 13*th European Conference*, Zurich, 6-12 September 2014, 596-611. https://doi.org/10.1007/978-3-319-10602-1_39

[10] Kong, Y., Gao, S., Sun, B. and Fu, Y. (2018) Action Prediction from Videos via Memorizing Hard-to-Predict Samples. Proc. 32*nd AAAI Conference on Artificial Intelligence*, New Orleans, 2-7 February 2018, 7000-7007.

[11] Wang, X., Hu, J., Lai, J., Zhang, J. and Zheng, W. (2019) Progressive Teacher-Student Learning for Early Action Prediction. *IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, 16-21 June 2019, 3556-3565.

[12] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1-32. https://doi.org/10.1162/neco.1997.9.1.1

[13] Graves, A. and Schmidhuber, J. (2005) Framewise Phoneme Classification with Bidirectional LSTM Networks. *IEEE International Joint Conference on Neural Networks*, Montreal, 31 July-4 August 2005, 846-853.

[14] Simonyan, K. and Zisserman, A. (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition. 3*rd International Conference on Learning Representations*, San Diego, 7-9 May 2015, 1-14.

[15] Cao, Z., Simon, T., Wei, S.E. and Sheikh, Y. (2017) Real-Time Multi-Person 2D Pose Estimation Using Part Affinity Fields. 30*th IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2016, 1302-1310. https://doi.org/10.1109/CVPR.2017.143

[16] Ryoo, M.S. (2011) Human Activity Prediction: Early Recognition of Ongoing Activities from Streaming Videos. 13*th International Conference on Computer Vision*, Barcelona, 6-13 November 2011, 1036-1043. https://doi.org/10.1109/ICCV.2011.6126349

[17] Soomro, K., Idrees, H. and Shah, M. (2016) Predicting the Where and What of Actors and Actions through Online Action Localization. *The IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 2648-2657. https://doi.org/10.1109/CVPR.2016.290

[18] Cao, Y., Barrett, D., Barbu, A., Narayanaswamy, S., Yu, H., Michaux, A., Lin, Y., Dickinson, S., Siskind, J.M. and Wang, S. (2013) Recognize Human Activities from Partially Observed Videos. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 1, 2658-2665. https://doi.org/10.1109/CVPR.2013.343