基于机器学习的手语识别系统研究与开发

尹昱杰　王启源　裴中正　施浩男　刘顺意

（北京信息科技大学，北京　100101）

摘　要：随着机器学习和计算机视觉领域的发展，计算机已经具有分析视频内容的能力。听障人士的交流困难主要是难以理解手语动作的含义，而借助计算机分析手语视频将其转化为文字，则可以解决以上交流难题。因此，提出基于机器学习的手语识别系统，先建立手语视频数据集，借助飞桨 AI Studio 平台使用时间分段网络（Temporal Segments Networks，TSN）算法框架进行相应的模型训练，再对手语视频进行抽帧，分析图像内容并对其进行预测，最后输出预测结果，从而实现对手语视频的文字转译。

关键词：机器学习；手语识别；时间分段网络（TSN）；模型训练

中图分类号：TP18；TP301.6　　**文献标识码：**A　　**文章编号：**1003-9767（2023）07-198-04

Research and Development of Sign Language Recognition System Based on Machine Learning

YIN Yujie, WANG Qiyuan, PEI Zhongzheng, SHI Haonan, LIU Shunyi

(Beijing Information Science and Technology University, Beijing 100101, China)

Abstract: With the development of machine learning and computer vision, computers can have the ability to analyze video content. The communication difficulties faced by hearing-impaired individuals are mainly due to the difficulty in understanding the meaning of sign language actions. By analyzing sign language videos with a computer and converting them into text, these communication difficulties can be solved. This paper proposes a sign language recognition system based on machine learning. Through the establishment of a sign language video dataset, the corresponding model training is carried out using the Temporal Segments Networks(TSN) algorithm framework with the help of the flying propeller AI Studio platform. It extracts frames from the sign language video, analyzes the image content and predicts it, and finally outputs the prediction results, so as to achieve the text translation of the sign language video.

Keywords: machine learning; sign language recognition; Temporal Segments Networks(TSN); model training

0　引言

在当今社会，听障人士是非常特殊的群体。不懂手语的人不能理解听障人士想表达的意思，而听障人士也不能表达自己的需求，导致交流困难。目前，许多手语识别项目存在使用不方便和准确率低等问题。为了使听障人士能够更方便快捷的交流，可以通过手机端将听障人士的手势语言转换为文字或语音，实现与他人的无障碍交流。

1　研究现状

目前，国内外卷积神经网络（Convolutional Neural Network，CNN）广泛应用于计算机视觉领域的识别和认知。在视频处理中，传统的二维卷积神经网络是对视频信息中的关键帧图像进行单帧图像信息的识别。但是，这样的识别方法缺少考虑时间维度的帧间运动信息，且手语信息大多是连续的手势动作，使用该方法会导致实时性不强。

收稿日期：2023-02-21

基金项目：北京信息科技大学大学生创新创业训练计划项目（项目编号：5112210832）。

作者简介：尹昱杰（2001—），男，广西桂林人，本科。研究方向：人工智能与机器学习。

在采样系统中，时间分段网络（Temporal Segments Networks，TSN）已经有了非常成熟的应用，由空间流卷积网络和时间流卷积网络构成。从整个视频中稀疏地采样一系列短片段的方式来代替稠密采样，既能捕获视频全局信息，又能去除冗余，减少计算量。目前，基于稀疏表示的视觉应用已经取得了大量研究成果 [1-3]。

在信息录入方面，部分研究者使用专门的摄像机完成，但是相对比较烦琐，不利于在手机上推广。基于此，衍生了一种 TLD（Tracking-Learning-Detection）算法，可用于处理视频出现的遮掩动作，及时完善跟踪 [4-6]。

2 算法实现

2.1 算法描述

手语识别系统需要对视频进行处理和分析，并确定最后的输出结果属于哪一个手语动作。因此，算法需要实现学习和推断两种功能 [7]。这里选择使用 TSN 算法，其计算公式为

$$TSN(T_1,T_2,\cdots,T_K)=H\{G[F(T_1,W),F(T_2,W),\cdots,F(T_K,W)]\} \quad (1)$$

该算法将视频分为多个片段，在每个片段中包括一帧图像和两个光流特征图，进行稀疏采样，并计算每个片段使用 TSN 中的 F 函数，得到一个分数，同时获得对分数的一个类别分布，再将这些分数使用 TSN 中的 g 融合均值函数进行计算得出结果，并对结果使用 H 函数进行概率计算，最后概率最高的便是该视频的类别 [8]。

2.2 算法过程

首先，将输入的视频进行数据预处理，再将处理后的帧集传入 TSN 网络。其次，利用片段一致性函数融合不同片段的分类分数，得到视频级预测 [9]。最后，将来自所有模态的预测融合，产生最终的预测 [10]。TSN 算法处理过程如图 1 所示。

2.3 数据预处理流程

首先，将传入视频进行分段，随机取帧，并形成帧集。其次，通过函数将帧集内的帧转化为实例对象，并将帧的实例对象进行裁剪和翻转。最后，将处理后的帧集数据归一化，保存输出给 TSN 网络。TSN 处理视频大致过程如图 2 所示。

3 实现和部署

3.1 数据集建立

本次使用 TSN 算法框架进行模型训练。先建立手语视频数据集，挑选比较常见的手语动作，分别为来（Come）、走（Walk）、请（Please）以及加油（Fighting）。另外，

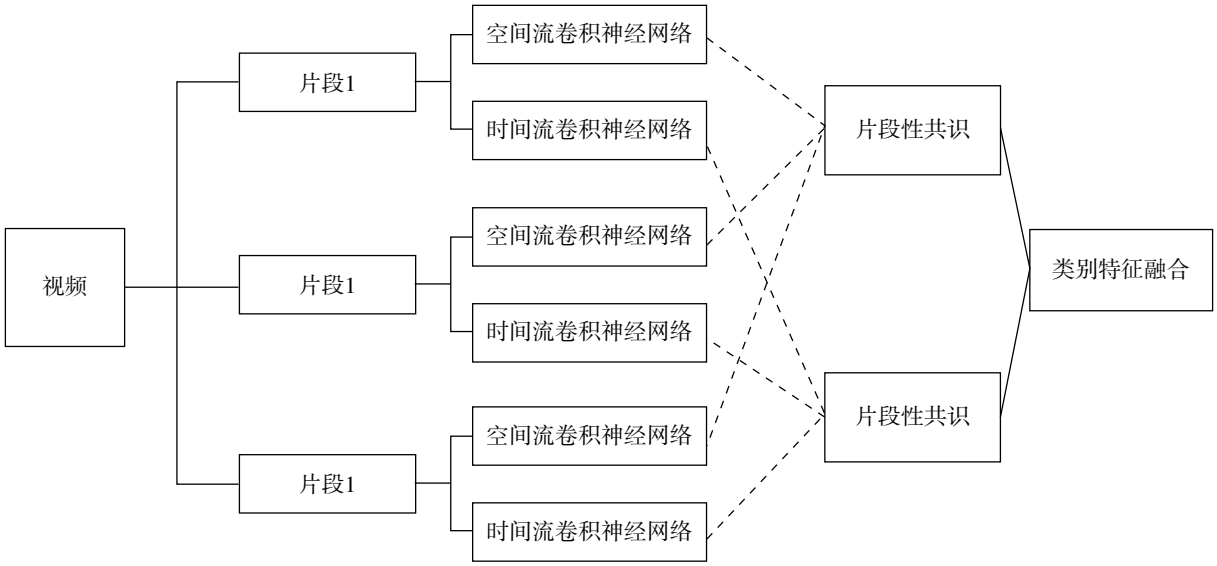

图 1 TSN 算法处理过程

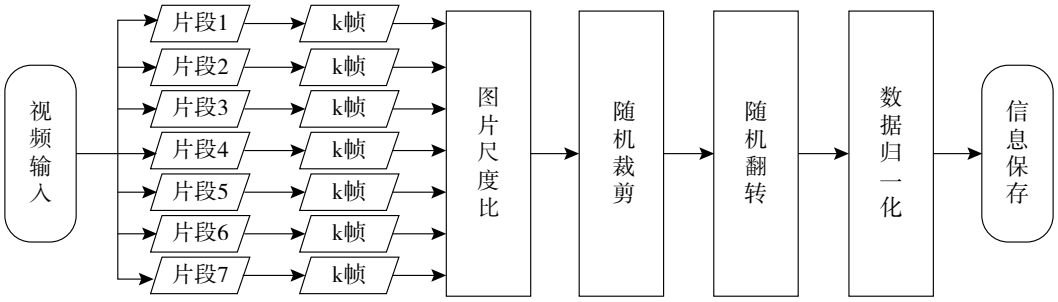

图 2 TSN 处理视频大致过程

特色的手语动作包括飞（Fly）和划桨（Paddle）。将手语视频按照对应的英文名称 + 数字的命名规则给每个视频标上标识，如 come0、walk2、please4 等。整理手语视频后，将视频按动作含义建立对应名称的文件夹，将这 6 个文件夹合并于存储视频集的文件夹，压缩并上传到 AI Studio 平台作为平台数据集，以便后序训练时导入。

3.2 模型部署

首先，将视频数据导入平台，借助 TSN 算法对手语视频进行抽帧。其次，将得到的图片信息存放在与对应视频同名的文件夹中，通过分析各图片信息生成相应 pkl 文件。最后，将这些 pkl 文件进行训练集、验证集、测试集的划分，验证集和测试集中各随机装有 6 个 pkl 文件，这 6 个文件分别来自 6 种不同的手语信息，训练集中装有剩余的 pkl 文件。

3.3 模型训练

在模型训练中，借助 AI Studio 平台提供免费的图形处理器（Graphics Processing Unit，GPU）Tesla V100 来提高模型训练效率。训练时，会在每一层卷积层对训练集的数据进行训练。对 valid 集的数据进行迭代分析，预测测试集的信息，并返回损失值、top_1_acc（最高预测率）、top_3_acc（排名前 3 的预测率）。其中，调参的文件 TSN.txt 内容如图 3 所示。

[MODEL]	[TRAIN]	[VALID]	[TEST]	[INFER]
name= "TSN" format = "pkl" num_classes= 10 seg_num = 3 seglen = 1 image_mean=[0.485,0.456,0.406] image_std= [0.229,0.224,0.225] num_layers = 50	epoch = 104 short_size = 240 target_size = 224 num_reader_threads = 1 buf_size = 1024 batch_size = 10 use_gpu = True num_gpus = 1 filelist="./dataset/train.txt" learning_rate=0.03 learning_rate_decay = 0.1 l2_weight_decay = 1e-4 momentum = 0.9 total_videos = 80	short_size=240 target_size=224 num_reader_threads = 1 buf_size=1024 batch_size = 2 filelist="./dataset/val.txt"	seg_num = 7 short_size = 240 target_size= 224 num_reader_threads = 1 buf_size = 1024 batch_size = 10 filelist="./dataset/test.txt"	short_size= 240 target_size= 224 num_reader_threads = 1 buf_size = 1024 batch_size = 1 filelist="./test/test.txt"

图 3 TSN.txt 中修改模型数据

将训练过程中每一层的损失值作为纵轴，卷积层数作为横轴，便于观察并及时调整模型参数。经过多次模型训练和调试，得到较为满意的训练结果，结果如图 4 所示。

3.4 实验结果和分析

训练过程中记录参数对模型识别准确率的影响，其中卷积层数的影响较大。多次尝试后，发现卷积层数为 104 时模型对手语视频的识别准确率最高。随机选择图片，使用模型对其进行预测，输出的预测结果和命中概率如表 1 所示。

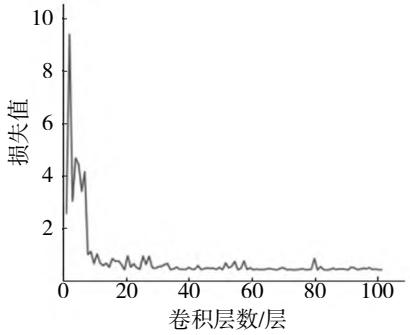

图 4 损失值和卷积层数的曲线分析

表 1 模型识别视频的结果与命中概率

视频名称	预测分类	命中概率 /%
walk10	Walk	98.70
fighting7	Fighting	99.93
come12	Come	92.30
fly7	Fly	99.97
paddle20	Paddle	99.90
please6	Please	96.90

3.5 模型对比

基于手语识别模型，参考不同算法模型计算 UCF101 大型动作识别数据集的动作识别准确率。将本文方法 TSN 和用于时间序列分析的 TSM 算法、使用两个独立的视觉通道进行视觉对象识别的双流卷积神经网络（two-stream）、用于视频分类和行为识别的密集三维卷积网络（Dense 3D Convolutional Networks，D3D）进行比较，不同算法在 UCF101 的准确率比较结果如表 2 所示。相较于数据量较小的手语数据集，TSN 算法模型能达到近似 94.9% 的准确率。

表 2 不同算法在 UCF101 的准确率

算法名称	识别准确率 /%
TSM	94.5
two-stream	92.5
D3D	97.6
TSN	94.9

4 结语

文章基于机器学习对手语视频识别进行开发和实践。从对 TSN 算法的学习和了解、手语视频数据集的建立、调参进行模型训练。结果表明，该模型具有较高的识别准确率。在未来的研究中，会继续完成手语数据集的扩展，进一步提高模型的准确率。

参考文献

[1] 韩雪平，吴甜甜．基于深度学习的人体行为识别算法 [J]. 数学的实践与认识，2019,49(24):133-139.

[2] TEODORO A M,BIOUCAS-DIAS J M,FIGUEIREDO M A T.Image restoration and reconstruction using targeted plug-and-play priors[EB/OL].(2021-03-09)[2023-02-15]. https://doi.org/10.1016/j.compeleceng.2021.107069.

[3] XU,S,ZHANG J,BO L,et al.Singular vector sparse reconstruction for image compression[J].Computers & Electrical Engineering,2021,91:107069.

[4] RODRÍGUEZ P,LARADJI I,DROUIN A,et al.Embedding propagation:smoother manifold for few-shot classification[C]//European Conference on Computer Vision,2020:121-138.

[5] YE H J,LI X C, ZHAN D C.Task cooperation for semi-supervised few-shot learning[C]// Proceedings of the AAAI Conference on Artificial Intelligence,2021:10682-10690.

[6] BARTSCH M V,MERKEL C,SCHOENFELD M A,et al.Attention expedites target selection by prioritizing the neural processing of distractor features[J]. Communications Biology,2021,4(1):814.

[7] 冯金源 . 基于深度神经网络的视频行为识别方法研究 [D]. 重庆 : 重庆大学 ,2021:23.

[8] 潘陈昕 . 深度学习在视频动作识别中的应用 [J]. 计算技术与自动化 ,2020,39(4):123-127.

[9] 戴兴雨 , 王卫民 , 梅家俊 . 基于深度学习的手语识别算法研究 [J]. 现代计算机 ,2021,27(29):63-69.

[10] 张珂 , 冯晓晗 , 郭玉荣 , 等 . 图像分类的深度卷积神经网络模型综述 [J]. 中国图象图形学报 ,2021,26(10):2305-2325.

（上接第 191 页）

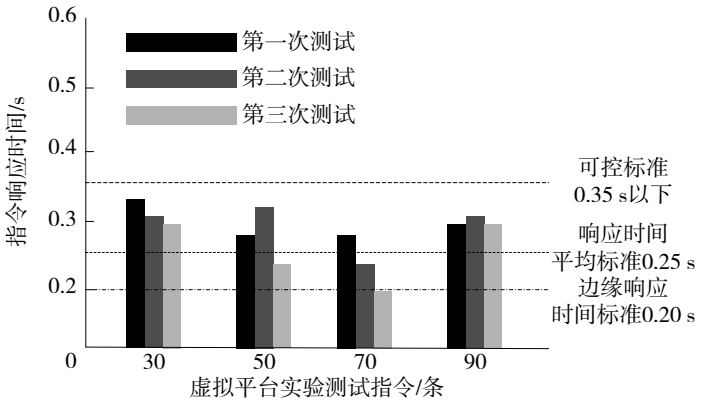

图 4　测试结果

和异常情况，增加实验处理过程中的安全性，可在降低实验误差的同时，便于相关人员进行资源共享、程序完善、远程操作等处理，能够营造一个仿真的实验环境，推动实验类平台进一步创新。

参考文献

[1] 张翔 . 基于虚拟实验平台的产品设计专业实践课程体系构建 : 以保定理工学院为例 [J]. 科幻画报 ,2022(12):164-165.

[2] 侯国栋 , 李媛 . 基于 " 智能 + 教育 " 的虚拟实验共享平台设计策略研究 [J]. 科技与创新 ,2022(11):12-15.

[3] 姚姗姗 , 董薇 , 拱溟昊 . 悬浮法生产可发性聚苯乙烯 (EPS) 虚拟仿真实验平台的设计与开发 [J]. 高分子通报 ,2022(1):86-90.

[4] 阎岳 . 基于 B/S 模式的实验室三维虚拟网络控制系统设计 [J]. 现代电子技术 ,2021,44(8):73-76.

[5] 陈立 . 基于 Unity 3D 和 AR 技术的虚拟实验室系统设计和仿真 [J]. 山西财经大学学报 ,2022,44(增刊 1):199-201.