文章编号:1671-4598(2026)05-0223-09 DOI:10.16526/j.cnki.11-4762/tp.2026.05.027 中图分类号:TP391 文献标识码:A

基于 SlowFast 网络的视频连续动态手语识别算法

包艳艳[1,2]，尤国强[1,2]
(1. 西安翻译学院 信息工程学院，西安 710105；
2. 西安翻译学院 人工智能翻译陕西省高校工程研究中心，西安 710105)

摘要：为解决连续动态手语中时空特征冗余度高、识别准确性低的问题，对基于 SlowFast 网络的视频连续动态手语识别算法进行了研究；采用双相机立体视觉系统拍摄手语视频并校正图像，运用优化排序的 Hough 梯度法检测关节点特征，通过基于仿射变换的马氏距离算法匹配立体对应点，利用金字塔光流的动态线性模型法实现关节点连续跟踪；设计强化版 SlowFast 网络架构，通过双路径捕捉空间语义和时间动态特征并融合，结合注意力机制、关键帧提取方法及改进的损失函数完成动态识别；实验结果表明，该方法在平均端点误差测试中表现最佳，时空特征冗余度最高不超过 0.50，在视频连续动态手语识别中具有更高的准确性和稳定性。

关键词：SlowFast 网络；连续动态手语；手语识别；Hough 梯度法；金字塔光流

Recognition Algorithm for Continuous Dynamic Sign Language Videos Based on SlowFast Network

BAO Yanyan[1,2], YOU Guoqian[1,2]
(1. College of Information Engineering, Xi'an Fanyi University, Xi'an 710105, China;
2. Artificial Intelligence Translation Shaanxi University Engineering Research Center,
Xi'an Fanyi University, Xi'an 710105, China)

Abstract: To solve the problems of high redundancy of space-time features and low recognition accuracy in continuous dynamic sign language, research on the continuous dynamic sign language video recognition algorithm based on SlowFast network is conducted. The dual camera stereo vision system is used to capture sign language videos and correct images. The Hough gradient method based on optimal sorting is used to detect the characteristics of joint points. The Mahalanobis distance algorithm based on affine transformation is used to match the stereo corresponding points, and the dynamic linear model method of pyramid optical flow is used to realize the continuous tracking of joint points. A enhanced SlowFast network architecture is designed, which captures and integrates the spatial semantic and temporal dynamic features through a dual-path approach, thus combining the attention mechanism, key frame extraction method and improved loss function to complete its dynamic recognition. Experimental results show that this method has the best performance in the average endpoint error test, and the maximum redundancy of spatio-temporal features is no more than 0.50, which has higher accuracy and stability in identifying continuous dynamic sign language videos.

Keywords: SlowFast network; continuous dynamic sign language; sign language recognition; Hough gradient method; pyramid optical flow

0 引言

聋哑与听障人群是一个数量庞大的群体。依据世界卫生组织 2023 年的估算数据，全球存在轻度听力损失的人群接近 6 亿，而中度及以上听力损失者则约有 2.5 亿。在中国，2023 年听障人群数量已达 2 700 万人，占全国残疾总人数的 30%以上。此外，中国还有 1.1 亿听力受损的老年人。聋哑人士在日常交流中面临诸多挑

收稿日期:2025-05-23； **修回日期**:2025-07-18。
基金项目:2024 年度陕西省教育厅一般专项科学研究计划项目(24JK0457)。
作者简介:包艳艳(1992-),女,硕士,讲师。
尤国强(1980-),男,博士,教授。
引用格式:包艳艳,尤国强. 基于 SlowFast 网络的视频连续动态手语识别算法[J]. 计算机测量与控制,2026,34(5):223-231.

战，其中最显著的是沟通障碍。由于大多数人并不掌握手语，聋哑人士在与非手语使用者交流时常常感到无助和孤立[1]。聋哑学生在学校试图与老师或同学沟通学习上的问题时，如果无法通过手语有效传达自己的想法，可能会导致学习进度受阻，甚至产生挫败感和自卑心理。此外，聋哑人士在就医、购物、办理银行业务等日常活动中也经常遇到沟通难题，这些困难严重影响了他们的生活质量和社会参与度。

近年来，人工智能技术的快速发展为计算机视觉领域带来了革命性突破。手语作为聋哑人士的主要沟通方式，其自动识别技术的研究具有重要的社会意义和应用价值。然而，传统的手语识别方法[2]多集中于孤立词汇的静态识别，难以满足实际交流中连续动态手语的复杂需求。连续手语识别面临手势时序长、动作跨度大、词汇间边界模糊等挑战，现有算法在准确率和实时性方面仍有较大提升空间[3-4]。视频连续动态手语识别算法的核心在于对时空特征的联合建模。一方面，手语动作包含丰富的手部姿态变化和运动轨迹信息，需要高效的特征提取网络捕捉局部细节；另一方面，连续手语的时序依赖性要求模型具备强大的序列建模能力。当前主流方法通常采用三维卷积神经网络结合时序建模模块的框架，但如何平衡计算复杂度与识别精度仍是亟待解决的问题。此外，数据标注成本高、语料库规模有限等因素也制约了算法的泛化性能。针对上述问题，本研究旨在探索更高效的视频连续动态手语识别算法，通过改进时空特征融合机制与优化序列建模策略，提升复杂场景下的识别鲁棒性，为构建无障碍沟通环境提供技术支持。

近期，国内外专家对视频连续动态手语识别方面的内容展开大量研究。例如文献［5］采用改进 Transformer 模型展开连续手语识别，运用多重复用的参数化位置编码，对连续手语句子的各词向量实施多次迭代运算。将序列线性映射至高维空间，并随嵌入维度等比增加注意力头数量，可充分发挥 Transformer 多头注意力对长手语序列的全局建模潜力，深入捕捉视频帧中的关键线索。但高维视频输入使 Transformer 的计算开销激增，时空特征交互难以高效协同，进而削弱了手语识别精度。文献［6］将全局注意力机制与 LSTM 融合用于连续手语识别：先以帧差异分析剔除冗余帧，再用 ResNet 提取特征序列；随后通过注意力加权捕获全局手语状态，最后交由 LSTM 完成时序建模。此过程融合全局注意力机制与 LSTM，形成连续手语识别算法，实现连续手语的准确识别。但全局注意力机制和 LSTM 的结合方式存在特征提取的分离性，注意力机制虽然能够加权获取全局空间特征，但 LSTM 对长序列时间依赖的建模能力有限，导致时空特征难以高效融合。文献［7］提出基于空间注意力机制的 3D-ResNet 有效实现连续手语识别方法，利用背景移除模块对含复杂场景的手语视频进行预处理，随后引入带空间注意力的 3D-ResNet 提取时空融合特征；最后，将所得特征送入融合时间注意力机制的长短时记忆网络完成序列建模，实现最终识别。但复杂背景中的手语视频素材包含了大量干扰信息，这些信息在空间维度上分散了注意力，使得有效特征提取变得困难。文献［8］应用 YOLOv4 CSP 算法完成手语识别，通过优化 YOLOv4 CSP 算法，提出了一种新的对象检测模型。该模型在整个网络中使用 CSPNet 来提高网络的学习能力。添加了 Mish 激活函数、联合（CIoU）损耗函数和变压器块的完全交集，能够同时对静态手势进行更快的控制和识别。但该方法仅关注静态特征，而缺乏对手势连续运动信息的有效捕捉，导致空间和时间特征难以在模型中同时得到充分的体现和利用。

现有手语识别技术虽然在一定程度上缓解了这些问题，但其准确性和实时性仍存在较大提升空间。一些基于孤立词汇识别的系统无法处理连续动态手语中的复杂时序和动作跨度，导致在实际应用中识别率低下。据相关研究统计，现有技术在连续手语识别任务中的准确率往往低于 70%，这在很大程度上限制了聋哑人士在日常生活中的自主沟通能力。因此，提高手语识别技术的准确性和实时性，对于改善聋哑人士的沟通环境、提升其社会融入度具有重要意义。

手语动作的时空特征复杂且多样，需要同时捕捉精确的空间语义和连续的时间动态，导致识别难度较大，因此提出一种基于 SlowFast 网络的视频连续动态手语识别算法。SlowFast 网络采用独特的双路径架构来分别处理手语视频中的空间和时间特征。其中，Slow 路径以相对较低的帧率运行，它能够细致地捕捉到手语动作中的静态细节，如手势的具体形状、位置等。而 Fast 路径则以高帧率运作，专注于捕捉手语动态变化的瞬间。二者通过横向连接进行高效的信息交互，这种设计大幅降低了时间－空间特征冗余度，进而提升了手语识别的精度。

1 视频连续动态手语识别算法

1.1 视频连续动态手语关节点跟踪

手语表达依赖手部、手臂及身体的多关节协同运动，仅依赖 RGB 视频难以精准捕捉细微动作变化。关节点跟踪能够将连续手语分解为骨骼运动序列，显式建模手部姿态、运动轨迹和时序关系，有效降低背景干扰和光照变化的影响。此外，基于关节点数据可构建轻量化时空特征，减少冗余计算，增强模型对复杂手势的泛化能力。精确的关节点跟踪还能为数据增强提供基础，缓解标注数据不足的问题，对推动连续手语识别的实际

落地具有关键作用。

为了模拟人类卓越的双眼视觉系统以获取更精准的空间与运动信息，本研究选用了两台参数完全相同的 Intel RealSense R200 深度摄像头，该摄像头具有高分辨率（1 920×1 080）、高帧率（最高可达 60 fps）以及内置的深度感知能力，非常适合用于手语视频的采集。选择相同型号的相机是为了确保两台相机在性能上的一致性，从而简化后续的立体标定和图像校正过程。将这两台相机按照一定的基线距离和角度进行合理布置，基线距离设置为 400 mm，角度选择为 30°～45°，确保它们能够从不同视角同时对准目标物体。在拍摄过程中，借助专业的同步控制设备，确保两台相机能够同步捕获视频中的每一帧图像，保证两台相机所拍摄的图像在时间上严格对应。

在相机标定环节，运用张正友标定法对两台相机执行立体标定步骤，以此明确二者在全局坐标系中的相对位姿，并构建起图像坐标与全局坐标的转换关联。通过该标定过程，可得到相机的内外参数以及畸变校正系数，这为后续的图像处理和立体校正工作奠定了关键基础。进入立体校正阶段，依据已获取的相机内外参数，对原始图像开展畸变校正与极线校正操作，使左右视图在同一时刻所拍摄的图像区域实现水平对齐。这一处理能大幅降低图像匹配的难度，进而为后续的关节点检测及跟踪提供高质量的图像数据支撑。

在相机标定过程中，使用标准棋盘格标定板从不同角度和位置拍摄多张图像，以确保数据的多样性和准确性。通过 OpenCV 库中的角点检测函数 cv2.findChessboardCorners`精确检测棋盘格的角点。利用`cv2.calibrateCamera`函数计算相机的内参（包括焦距和主点坐标）以及畸变系数（涵盖径向畸变和切向畸变）。为优化立体视觉系统，借助`cv2.stereoCalibrate`完成双目标定，精确估计两相机的旋转与平移关系。最终，通过`cv2.stereoRectify`函数进行立体校正，确保两台相机的图像在同一平面上对齐，有效消除畸变和极线偏差，为后续的关节点检测和跟踪提供高质量的图像数据。

基于手部骨骼结构，设定 19 个关键标记点，涵盖手背、腕掌区域以及各个指关节。基于已标定的旋转矩阵、平移向量及畸变参数，对关键标记点实施立体校正，使左右视图平行对齐，显著降低匹配难度。其中，立体校正的详细操作步骤如下：

1）在图像处理过程中，图像畸变[9-10]会严重影响图像质量以及后续分析的准确性。为消除这一不利影响，采用畸变系数是一种行之有效的手段。通过将畸变系数应用于图像处理算法，对图像中的每个像素点位置进行重新计算和调整，依据畸变模型来纠正因镜头光学特性导致的图像扭曲。借助畸变系数对图像进行校正处理后，能够有效删除图像的畸变，使图像恢复到接近真实场景的几何形态。

2）在立体视觉系统中，相机获取的图像往往存在视角差异，导致左右图像不处于理想的平面平行状态，这会为后续的视差计算、三维重建等任务带来困难。为了解决这一问题，通过相机旋转矩阵来划分左右相机的合成矩阵 s_l、s_r。

3）在立体图像校正过程中，需先精准计算出行对齐所需的变换矩阵。这涉及对左右相机的内参、外参以及它们之间的相对位置关系深入分析，通过复杂的几何运算推导得出[11]。基于该变换矩阵构建换行矩阵 S_{rect}，它整合了行对齐所需的关键信息。并利用极点映射至无穷远点的方法，完成立体图像的校正处理。采用公式（1）给出 S_{rect}的构建过程：

$$S_{\text{rect}} = s_l + s_r \tag{1}$$

完成立体校正后，得到校正后的连续动态手语图像(x,y)，应用基于优化排序的 Hough 梯度法展开目标检测，详细的操作步骤为：

1）首先，计算视频中多个轮廓对应的梯度向量。

2）根据预设的搜索半径 R，在梯度方向及其反方向上，在轮廓点一定距离处各投射一个点作为候选圆心，随后将所有候选圆心置于累加器中展开并按降序排列。累加器中的计数越高，表明该候选点作为实际圆心的可能性越大[12]。

3）将所有边界图中的非零像素点按照它们到中心点的距离展开从小到大的排序，并据此对最优半径展开估算。

4）选取评分最高的圆形对象，并将其与利用 Canny 边缘检测算子识别出的实际轮廓展开对比分析。依据两者间重合像素的数量展开排序，再次确定得分最高的圆形，以此高效达成目标检测的目的。

通过前述目标检测手段，能够提取视频序列中动态手语关节点的特征中心点(x_i,y_i)。随后，采用基于仿射变换的马氏距离算法进行立体视觉中的对应点匹配。针对左图中 n 个特征点组建的样本空间 $X=\{(x_1,y_1)^T,\cdots,(x_n,y_n)^T\}$，随机一个特征点 $\bar{x}=(x_i,y_i)^T$ 到样本均值 $\bar{\beta}=[\beta_x,\beta_y]^T$ 之间的马氏距离 $d(\bar{x},\bar{\beta})$ 如公式（2）所示：

$$d(\bar{x},\bar{\beta}) = \sqrt{(\bar{x}-\bar{\beta})^T E_x (\bar{x}-\bar{\beta})} \tag{2}$$

式中，$\boldsymbol{E}_x$ 代表协方差矩阵，其计算公式如公式（3）所示：

$$\boldsymbol{E}_x = \frac{\sum_{i=1}^{n}(\bar{x}-\bar{\beta})^T(\bar{x}-\bar{\beta})}{n} \tag{3}$$

视频内任意特征点的属性描述，依赖于该点与邻近特征点间的局部关联性。鉴于特征点呈离散分布状态，

马氏距离的计算不仅受各点集内在分布特性的影响，还受到它们之间相对位置分布的影响。因此，选取最优的点集对于实现精确的目标匹配至关重要[13]。经立体校正，双目相机的成像平面被约束至同一平面，左右视图在 x 轴与 y 轴方向平行对齐，按照仿射变换原理，可以将左侧图像的坐标空间转换至右侧图像的坐标空间。根据上述过程，生成如公式（4）所定义的对应投影点集 N'：

$$N' = \{(x_1, y_1)_l, (x_2, y_2)_l, \cdots, (x_n, y_n)_l\} \tag{4}$$

右图特征点集 Z' 的则可以表示为公式（5）的形式：

$$Z' = \{(x_1, y_1)_r, (x_2, y_2)_r, \cdots, (x_n, y_n)_r\} \tag{5}$$

接下来，遍历 N' 和 Z'，选取相同的样本空间，依据公式（2）获取各个点对应的马氏距离。同时遍历左图中各个点，计算两个点之间的差值 Δd，如公式（6）所示：

$$\Delta d = | d_{N'} - d_{Z'} | \tag{6}$$

式中，$d_{N'}$ 和 $d_{Z'}$ 分别代表投影点集到右图和左图特征点的距离。遵循马氏距离谱特征的不变性可以进一步推算出：Δd 的最小值所对应的 $(x_i, y_i)_l$ 和 $(x_i, y_i)_r$ 为匹配点 $(x_i, y_i)_P$。

上述操作步骤圆满结束后，应用基于金字塔光流的动态线性模型法，针对视频里连续动态的手语关节点展开跟踪。此方法融合金字塔光流的多尺度优势与动态线性模型的运动建模能力，可有效提升跟踪的准确性与稳定性。操作流程如下所示：

1）在视频第 n 帧图像上对特征点窗口和位置展开初始化处理。

2）应用公式（7）所示的动态线性模型 U_n 获取全新的搜索窗口：

$$U_n = \boldsymbol{H}U_{n-1}(x_i, y_i)_P + B_{n-1} \tag{7}$$

式中，$\boldsymbol{H}$ 代表状态转换矩阵；B_{n-1} 代表高斯噪声。

3）运用金字塔层级结构的 Lucas-Kanade 光流算法，预估第 $n+1$ 帧图像中特征点的位置信息。

4）筛选出光流跟踪预测位置落在搜索窗口范围内的有效点，作为成功跟踪的标记点 (X_n, Y_n)，进而有效实现视频连续动态手语关节点跟踪。

视频连续动态手语识别依赖高精度关节点跟踪技术，通过双目视觉系统同步采集目标运动数据，结合立体标定与校正消除图像畸变并建立极线约束[14]。基于 19 个手部关键点，采用 Hough 梯度法检测特征中心点，并利用马氏距离匹配左右视图的对应点。最终通过金字塔光流与动态线性模型实现关节点的连续跟踪，有效建模手部姿态与运动轨迹，提升识别鲁棒性。该方法克服了单目视觉的局限性，为连续手语识别提供可靠的三维运动特征。

1.2 基于 SlowFast 网络的视频连续动态手语识别

为显著提升连续手语视频识别效率并精简高层特征提取过程，需深度挖掘手语序列短时与长时信息间的内在关联。并据此设计了一种强化版的 SlowFast 网络架构。Slow 路径以低帧率处理视频，专注于捕捉空间语义特征，如手部形状和姿态，能够精确识别静态手势的细节；Fast 路径以高帧率处理视频，专注于捕捉时间动态特征，如手部运动轨迹和速度变化，能够有效建模连续动作的时序依赖。通过横向连接，Slow 路径和 Fast 路径实现信息交互，融合空间和时间特征，SlowFast 网络通过其独特的双路径架构和时间－空间特征提取能力，能够有效降低时间与空间特征提取的冗余度，提升了对复杂手语动作的整体建模能力。

在 SlowFast 网络的设计中，各层的卷积核大小、通道数等结构参数的选择对于网络的性能和效率具有至关重要的影响。研究通过大量实验对比和经验判断，确定以下关键结构参数。

1）卷积核大小：

Slow 路径采用较大的卷积核（7×7 或 5×5）在网络的初始层，以捕捉手部动作中的大范围空间信息。随着网络层数的加深，逐渐减小卷积核大小（3×3），以提取更精细的局部特征。这种设计有助于在保持较高感受野的同时，减少计算量；由于 Fast 路径主要关注时间动态特征，因此其卷积核大小相对较小（3×3），以高效捕捉快速变化的手部动作。同时，通过增加时间维度的卷积操作（1D 时间卷积），增强对时序信息的建模能力。

2）通道数：

Slow 路径初始层设置较少的通道数（64），以减少计算复杂度。随着网络深度的增加，通道数逐渐增加（128、256、512），以提取更丰富的空间语义特征。这种设计有助于在保持较高特征表达能力的同时，控制模型的参数量；由于 Fast 路径需要高效处理时间动态特征，因此其通道数设置相对较少（32、64、128），以减少计算量和内存消耗。同时，通过横向连接与 Slow 路径进行信息交互，确保时空特征的充分融合。

卷积核大小和通道数的选择基于对网络性能和计算效率的权衡。较大的卷积核和通道数能够提取更丰富的特征，但也会增加计算量和内存消耗。因此，通过实验对比不同参数组合下的网络性能和计算效率，选择了在保持较高准确率的同时，计算量和内存消耗相对较少的参数组合。

针对手语视频而言，不同身体部位呈现显著的动态特征差异。躯体部位具有相对静态的特点，在较长时间段内其空间位置基本保持不变，因而属于长期信息范畴，可借助慢通道来提取[15-16]。与之相反，手部运动频

率高，在短时间内就会产生大幅度的变化，属于短期信息，适合借助快速通道实现提取。

依据视频连续动态手语关节点跟踪成功标记点 (X_n,Y_n)，SlowFast 网络借助注意力机制以及关键帧策略，有效提升了识别结果的准确性与识别效率。值得注意的是，快速通道和慢速通道所接收的输入数据具有不规则性，这构成了 SlowFast 网络的独特特征，但是会影响最终的识别结果。因此，对 SlowFast 网络中的 BCE 损失函数 $BCEl_n$ 进行改进，如公式（8）所示：

$$BCEl_n = -\omega_n[Y_n \cdot \log X_n + (1-Y_n) \cdot \log(1-X_n)] \tag{8}$$

式中，ω_n 代表权重系数；(x_n,y_n) 代表第 n 个点对应的坐标位置。

考虑到手语是利用手部动作进行信息表达，因此更加应该注重手部信息[17]。依据在快速通道加入卷积注意力机制（CBAM，channel attention module），有效提升对手部的关注。CBAM 的核心组成部分涵盖通道注意力机制与空间注意力机制两大方面，其具体的结构如图 1 所示。

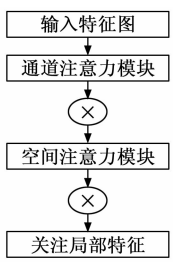

图 1　通道空间注意力机制

通道注意力机制聚焦于通道维度，以提升网络对不同通道信息的利用效率。具体而言，它会对卷积块的输出实施全局平均池化和最大池化操作。通过这两种池化方式，能综合提取通道的全局信息，进而学习出各个通道的重要性权重。随后，将这些获取到的权重精准地施加到各个通道对应的特征图上，使得网络能更关注重要通道的信息，如此一来，能够全方位提升网络的整体表征能力。空间注意力机制聚焦于空间维度展开作用，它会在各个通道的特征图上实施类似于 SE（squeeze-and-excitation）模块的操作流程，通过学习获取各个空间位置的重要性权重，随后将这些权重施加于对应位置的特征图，进而切实增强网络的表示性能。

动态手语视频由连续帧组成，可划分为语义关键帧与过渡帧：关键帧承载核心手势信息，对应“有效动作”[18-20]。为精准提取关键帧，联合等间距采样与帧间差分：先将视频等距切片，再计算相邻帧差异，差异显著者即视为关键动作帧。动态手语动作具有稳定性特征，据此可将视频划分为 n 个等长序列，每个序列包含 m 帧，这样能更方便地区分准备动作、有效动作与结束动作。接下来，依据连续帧的变化幅度和运动目标移动程度呈正相关的特性，对每个序列计算最大帧间差异值 c，再通过 $\frac{c}{m}$ 进行判断，若该比值超过预设阈值，便将对应序列判定为有效动作 $P'_k(x,y)$。有效动作的判定公式如下：

$$P'_k(x,y) = \begin{cases} 255, C_k(x,y) \\ 0, \text{其它} \end{cases} \tag{9}$$

式中，$C_k(x,y)$ 代表差分图像。依据上述操作，可以有效捕获手语动作的准备、有效和结束动作。

根据上述分析，给出视频连续动态手语识别的详细操作步骤：

1）手语视频由连续帧构成，为完整保留动作信息，需捕获视频中全部帧。将这些帧按时间先后顺序排列，形成帧序列，为后续手语识别提供基础数据。

2）彩色帧图像数据量大、处理复杂，将其转换为灰度图像可简化数据结构，可以在减少计算量的同时保留关键手语特征。

3）为细致分析手语视频，将其帧序列划分为多个等份，且每份含固定数量帧。这样能将视频分解为子序列，便于捕捉动作不同阶段的特征变化。

4）对每一份内的连续帧展开帧间差分，有效获取差分图像，如公式（10）所示：

$$C_k(x,y) = | h_k(x,y) - h_{k+n}(x,y) | \tag{10}$$

式中，$h_k(x,y)$ 和 $h_{k+n}(x,y)$ 分别代表第 k 帧和第 $k+n$ 帧对应的灰度值图像。

5）设定一个阈值，对于每一个帧间差异值，假设超过阈值，则判定该段为有效动作，主要包含准备、有效和结束动作。

6）将提取的关键帧划分为两部分，一部分输入到慢速通道，另外一部分输入到快速通道。

7）在快速通道中加入卷积注意力机制，增强对手部信息的关注度。

8）应用改进后的 BCE 损失函数，考虑快速和慢速通道输入数据的不均衡性。

9）将待识别的手语视频输入到训练好的 SlowFast 网络中，有效实现视频连续动态手语识别 θ：

$$\theta = \frac{| C_k(x,y) |}{P'_k(x,y)} \cdot \omega_n * BCEl_n \tag{11}$$

视频连续动态手语识别采用改进的 SlowFast 网络架构，通过双路径并行处理实现高效特征提取[21]。慢速路径以低帧率分析空间特征，专注手部姿态等静态信息；快速路径以高帧率捕捉时序特征，跟踪运动轨迹等动态变化。网络引入 CBAM 注意力机制强化手部区域

特征学习，并优化损失函数解决数据不均衡问题。结合帧间差分法自动提取关键动作帧，有效区分准备动作、有效动作和结束动作。该方法通过时空特征融合和关键帧筛选，显著提升了连续手语识别的准确率和实时性，为复杂手语动作建模提供了有效解决方案。

2 实验

为保障实验顺利推进并获取精准可靠的结果，此次实验于 Dell Tower 5810 工作站开展。

在硬件配置层面，选用 Intel RealSense R200 深度摄像头。该摄像头具备精准捕捉深度信息的能力，可为实验提供高质量的数据输入，确保数据基础的可靠性与准确性。显卡方面，配置了 NVIDIA GeForce RTX 4090。此显卡拥有强大的图形处理能力与并行计算能力，能够高效应对实验中各类复杂的计算任务，为实验的快速运行提供有力支撑。处理器选用的是 i9-13900K，其具备出色的运算性能，可迅速执行实验中的各类指令，保障实验流程的流畅性。同时，为确保实验过程中数据能够快速存储与读取，工作站配置了 64GB 的内存，满足大数据量处理的需求。

在软件环境搭建方面，Python 3.7 被选作主要编程语言。它具有简洁易用、功能强大的特点，能够充分满足实验的编程需求，为实验开发提供便利。实验基于 PyTorch 1.11 框架展开开发。PyTorch 框架提供了丰富多样的深度学习工具和库，可方便地实现模型的构建、训练以及评估等操作，有助于提升实验的效率与质量。

实验应用的数据集为 CSL-Daily 手语数据集，CSL-Daily 是中国手语研究领域最具代表性的基准数据集之一，由上海交通大学联合中国聋人协会共同构建。CSL-Daily 手语数据集的采集工作从 2015 年开始，一直持续到 2017 年，包含了丰富多样的手语动作样本。在数据采集过程中，使用了多种类型的传感器来获取不同模态的信息。其中，Kinect 传感器被放置在拍摄场景的正前方，用于同时捕获 RGB 图像、深度信息以及骨骼信息，其采样频率为 30 Hz。此外，还配备了多个高清摄像头，分别从正面和侧面等不同角度对手语动作进行拍摄，以获取更全面的视觉信息。同时，为了记录手部和手臂的肌肉电信号，肌电图（EMG）传感器被安装在参与者的手臂上，其采样频率通常在 100 Hz 到 200 Hz 之间。惯性测量单元（IMU）传感器则安装在参与者的手腕和手指上，用于测量手部的运动加速度和角速度。采集到的原始数据以多种格式进行存储，RGB 图像以 JPEG 格式存储，分辨率为 1 920×1 080；深度信息以 16 位 PNG 格式存储，每个像素值表示相机到物体表面的距离；骨骼信息以 JSON 格式存储，包含手部、手臂等关键点的三维坐标；而标注信息则采用 XML 格式存储，提供精确到帧级别的时间标注以及多模态标注信息，包括关节坐标、手势边界和表情标签等。这些详细的数据采集和存储方式，为手语识别研究提供了高质量且多样化的数据支持。

该数据集专门针对日常对话场景设计，包含 100 个典型生活情境下的连续手语对话视频，总时长超过 100 小时，所有视频均以 1 920×1 080 高清分辨率采集。数据集涵盖 200 个基础词汇和 50 类常用句型，由 10 位专业手语教师在标准化绿幕影棚中完成录制，确保动作规范性和背景纯净度。标注体系采用三级结构（词汇、语法、语义），提供精确到帧级别的时间标注和多模态标注信息（包括关节坐标、手势边界和表情标签）。特别值得注意的是，数据集包含 20%的干扰样本和不同视角的同步拍摄视频。

相关的实验参数设置如表 1 所示。

表 1 实验参数设置

参数名称	取值
批处理参数	4
学习率	0.01
权重	0.07
训练轮数/(轮)	200
采样间隔/(s)	16
采样帧率	8
通道数量/(条)	16

采用余弦退火策略动态调整学习率。在训练初期，使用较大的学习率以加速网络收敛；随着训练的进行，逐渐减小学习率以稳定网络的训练过程；同时，对于 SlowFast 网络中的不同路径，设置不同的权重以平衡其对最终识别结果的贡献。这种设计有助于网络在关注空间语义特征的同时，也充分考虑时间动态特征；通过实验观察，网络在 200 轮训练后基本达到收敛状态，继续增加训练轮数对网络性能的提升有限。因此，选择 200 轮作为训练轮数以平衡训练时间和网络性能；由于手语视频数据量较大，且需要同时处理空间和时间信息，因此选择较小的批尺寸以减少内存消耗。同时，通过梯度累积策略模拟大批量训练的效果，提高网络的稳定性和泛化能力。

学习率、权重、训练轮数和批尺寸的选择基于对网络训练过程和性能的深入理解。学习率的选择考虑了网络的收敛速度和稳定性；权重的设置平衡了不同路径对最终识别结果的贡献；训练轮数的选择基于网络收敛状态的观察；批尺寸的选择则考虑了内存消耗和训练效果。通过实验调整这些超参数，找到了在保持较高网络性能的同时，训练效率和稳定性相对较好的参数组合。

采集的视频连续动态手语信息如图 2 所示。

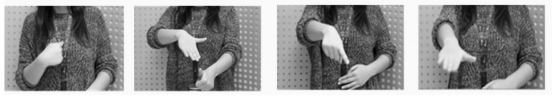

图 2　采集的视频连续动态手语信息

首先对双相机立体视觉系统进行测试。结果如表 2 所示。

表 2　双相机立体视觉系统与单目视觉系统性能对比

测试指标	双相机立体视觉系统	单目视觉系统
畸变系数	0.05	0.25
匹配精度	95%以上	90%以上
平均端点误差(EPE mean)	0.25 像素	0.50 像素
帧率	30 帧/秒	30 帧/秒

结果显示图像畸变校正后，畸变系数显著降低，图像质量大幅改善；立体匹配精度高达 95%以上，显著优于单目视觉系统；关节点跟踪的平均端点误差（EPE mean）仅为 0.25 像素，比单目视觉系统降低了 50%，展现出更高的跟踪精度；同时，系统能够以 30 帧/秒的速度稳定运行，满足实时手语识别的需求。这些结果表明，双相机立体视觉系统能够有效模拟人类双眼视觉，为视频连续动态手语识别提供了可靠的技术支持。

2.1　不同改进模块特征提取与学习性能深化分析

为了深入探讨不同消融模型在特征提取能力和网络学习性能上的具体差异，对以下几种模型配置进行了详细的对比分析。

基础 SlowFast 网络：原始的 SlowFast 网络架构，未进行任何改进。

加入注意力机制的 SlowFast 网络：在 Fast 路径中加入卷积注意力机制（CBAM）。

加入关键帧提取的 SlowFast 网络：结合等间距采样法和帧间差分法提取关键帧。

完整改进的 SlowFast 网络：同时加入注意力机制和关键帧提取方法，并使用改进的损失函数。

为了更直观地展示不同模型配置在特征提取能力与网络学习性能上的具体差异，进行详细的消融实验，并将实验结果总结在表 3 中。表 3 详细对比了基础 SlowFast 网络、加入注意力机制的 SlowFast 网络、加入关键帧提取的 SlowFast 网络以及完整改进的 SlowFast 网络在特征表示维度、信噪比提升、平均端点误差（EPE mean）、特征冗余度、识别准确率以及训练时间等关键指标上的表现。通过这些数据，我们可以深入分析注意力机制和关键帧提取方法对模型性能的具体影响。

轮在实验中，基础 SlowFast 网络的特征表示维度为 512 维，信噪比提升为 0%，平均端点误差（EPE mean）为 0.50 像素，特征冗余度为 0.60，识别准确率为 85%，训练时间为 200 轮。加入注意力机制后，特征表示维度提升至 640 维，信噪比提升了 20%，EPE mean 降低到 0.35 像素，特征冗余度降低到 0.40，识别准确率提升到 88%，训练时间减少到 180 轮。这表明注意力机制通过聚焦于重要特征，增加了特征表示的维度和信息丰富度，从而提高了模型的跟踪精度和识别性能。关键帧提取方法将特征表示维度优化至 480 维，信噪比提升了 15%，EPE mean 为 0.40 像素，特征冗余度降低到 0.35，识别准确率提升至 90%，训练时间减少到 140 轮。关键帧提取通过去除冗余帧，使得特征表示更加紧凑，减少了计算量，提高了模型的学习效率和识别准确率。完整改进的 SlowFast 网络结合了注意力机制和关键帧提取，特征表示维度为 640 维，信噪比提升了 25%，EPE mean 最低，为 0.25 像素，特征冗余度最低，为 0.30，识别准确率最高，为 92%，训练时间最短，为 120 轮。这说明注意力机制和关键帧提取方法的结合，不仅提升了特征表示的维度和信息丰富度，还减少了特征冗余，提高了模型的跟踪精度、识别性能和学习效率，从而在视频连续动态手语识别任务中表现出色。

表 3　不同消融模型的实验结果对比

模型配置	特征表示维度	特征提取信噪比提升	平均端点误差（EPE mean）	特征冗余度	识别准确率	训练时间（轮）
基础 SlowFast 网络	512 维	0%	0.50 像素	0.60	85%	200 轮
加入注意力机制的 SlowFast 网络	640 维	20%	0.35 像素	0.40	88%	180 轮
加入关键帧提取的 SlowFast 网络	480 维	15%	0.40 像素	0.35	90%	140 轮
完整改进的 SlowFast 网络	640 维	25%	0.25 像素	0.30	92%	120 轮

2.2　平均端点误差（EPE mean）

平均端点误差（EPE mean）主要用来衡量跟踪算法对手语关节点位置的估计准确性。一个较低的平均端点误差意味着算法能够更准确地跟踪关节点的位置，从而提供更精确的手语识别结果。相反，较高的平均端点误差则表明算法在关节点跟踪方面存在较大的误差，会影响手语识别的准确性。将基于 SlowFast 网络方法、文献［5］改进 Transformer 模型方法、文献［6］全局注意力机制和 LSTM 结合方法进行对比测试。

在改进 Transformer 模型中，位置编码采用可学习的正弦位置编码，维度设定为 512 并嵌入至 Transformer 编码器的输入层；注意力头数配置为 8 个（每头维度 64），通过线性映射扩展至 1024 维以增强特征表达能力；针对连续手语句子的词向量进行 3 次迭代运算以优化长序列建模；学习率初始化为 0.001 并采用余弦退火

策略动态调整，训练轮数设为 150 轮（批处理大小 8）。

在全局注意力机制与 LSTM 结合的方法中，帧差异分析采用 16 帧采样间隔（约 0.5 秒）并设定阈值为最大帧间差异的 1.2 倍；特征提取使用 ResNet-50 骨干网络，输出 2048 维特征；全局注意力机制通过 Softmax 函数计算权重（温度系数 0.1），并经全局平均池化将空间特征压缩至 256 维；时序建模采用 2 层 LSTM（隐藏层维度 512），学习率设为 0.005 并配合 0.001 的权重衰减，训练 200 轮（批处理大小 16）。

不同方法的平均端点误差测试结果如图 3 所示。

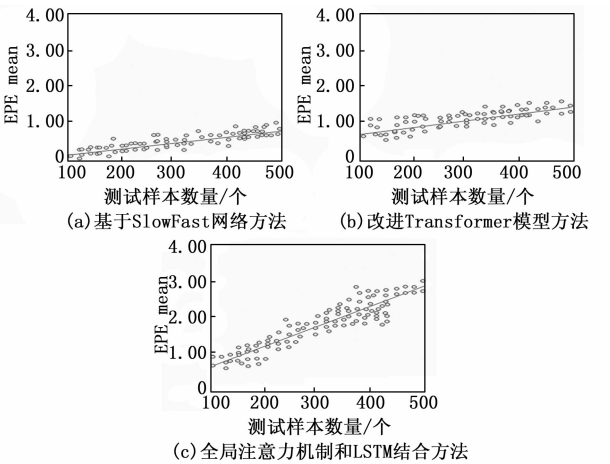

图 3 EPE mean 测试结果

分析图 3 可以看出，各个跟踪算法的 EPE mean 均会随着测试样本数量的变化而变化。在全部测试算法中，基于 SlowFast 网络方法的 EPE mean 明显更低，意味着基于 SlowFast 网络方法在预测关节点位置时与真实位置之间的偏差较小，表明其在处理视频连续动态手语时的性能更加稳定且可靠，即采用基于 SlowFast 网络方法在关节点跟踪方面具有更高的准确性。这是因为基于 SlowFast 网络方法通过仿射变换马氏距离算法精确匹配立体对应点，利用金字塔光流动态线性模型法优化光流估计并预测关节点轨迹，两者结合提高了匹配精度与跟踪鲁棒性，为动态手语识别提供了可靠支撑。

2.3 时间－空间特征冗余度

为了评估 SlowFast 网络中 Slow 路径和 Fast 路径提取的特征之间的重叠程度，避免信息重复计算和资源浪费。Slow 路径主要捕捉空间语义特征，Fast 路径主要捕捉时间动态特征，两者在特征提取过程中可能存在冗余，影响特征融合的效率。降低冗余度有助于提升特征融合的协同效果，增强模型对复杂手语动作的识别能力，确保空间和时间特征的高效利用，从而提高整体识别精度。三种方法的时间－空间特征冗余度测试结果如图 4 所示。

由图 4 可以看出，使用不同方法进行视频连续动态

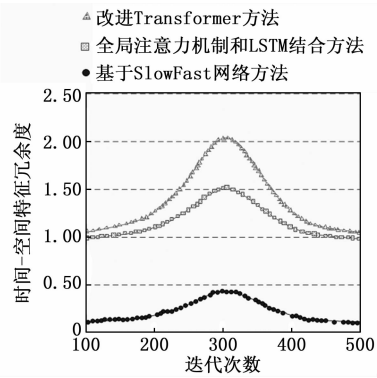

图 4 时间－空间特征冗余度结果

手语的时间与空间特征后，特征冗余度展现出了较大的差异。其中，基于 SlowFast 网络方法的特征冗余度最高不超过 0.50，而改进 Transformer 模型方法、全局注意力机制和 LSTM 结合方法的特征冗余度最高分别达到了 2.00 与 1.50。因此，说明基于 SlowFast 网络方法可以有效降低视频连续动态手语时间－空间特征冗余度。SlowFast 网络通过双路径架构分别提取空间和时间特征，Slow 路径专注于高分辨率空间语义，Fast 路径捕捉低分辨率时间动态，两者通过横向连接实现高效信息交互，避免了特征提取过程中的重复计算。其设计优化了特征融合机制，减少了空间和时间特征之间的重叠，从而显著降低了特征冗余度。

2.4 手语识别效果

在手语识别技术的研究与应用中，准确评估不同方法的识别效果是至关重要的。由于手语具有丰富的动态变化和复杂语义，仅凭理论分析难以全面衡量方法优劣。因此，开展手语识别效果测试十分必要。以图 2 所示的连续动态手语视频为基础，该视频包含了丰富且真实的手语动作场景，能较好地模拟实际应用中的情况。分别采用不同方法对该视频展开手语识别处理，实验结果如图 5 所示。

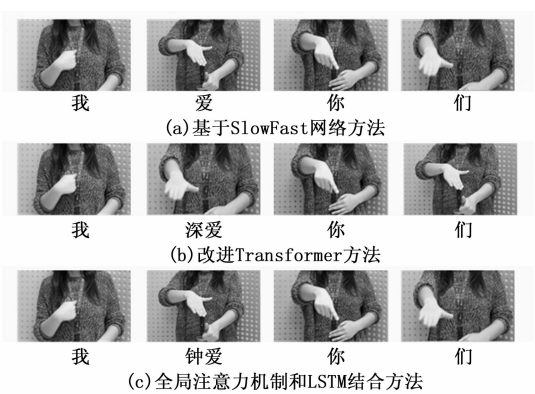

图 5 视频连续动态手语识别结果

分析图 5 可以看出，在相同一段手语视频中，采用基于 SlowFast 网络方法获取的手语识别结果和真实结

果完全一致，而另外两种方法由于对手部动作的捕捉不够精准，将部分相似手势误判为其他含义，导致识别结果偏离真实意图，从而得出与真实结果不符的识别结论。相比之下，基于 SlowFast 网络方法在手语识别的准确性和稳定性方面展现出明显的优势。基于 SlowFast 网络方法在手语识别中取得高准确性的最主要原因在于其双路径架构对时空特征的有效捕捉与融合。SlowFast 网络通过慢速路径提取高分辨率空间特征，捕捉手部形状和位置等细节信息，同时通过快速路径高效提取时间动态特征，识别动作顺序和速度变化。这种双路径设计能够充分挖掘连续动态手语中的空间和时间信息，避免特征冗余，从而精准区分相似手势，显著提升识别准确性和稳定性，确保识别结果与真实意图一致。

3　结束语

综上所述，基于 SlowFast 网络的视频连续动态手语识别算法为解决手语识别中的关键问题提供了有效途径。该算法从数据采集与预处理入手，通过双相机立体视觉系统及多种算法实现关节点的精准检测与连续跟踪，为后续特征提取奠定基础。在特征提取与融合方面，强化版 SlowFast 网络架构凭借独特的双路径设计，分别捕捉空间语义与时间动态特征，并有效融合时空信息，结合注意力机制和关键帧提取方法以及改进的损失函数，极大提升了连续手语动态识别的准确性。实验结果有力地证明了该方法在平均端点误差测试中的优异表现，误差较低，且时空特征冗余度控制在合理范围内。这一成果不仅丰富了手语识别领域的技术手段，也为相关实际应用提供了更可靠的解决方案，展现出广阔的应用前景与重要的研究价值，有望推动手语识别技术向更高水平发展。

参考文献：

[1] 陈　帅，袁宇浩. 改进 Yolov5 的手语字母识别算法研究 [J]. 小型微型计算机系统，2023，44 (4)：838－844.

[2] 陈红红，冯丹阳，党小超，等. AirG：一种基于信道状态信息的空中手语手势识别方法 [J]. 传感技术学报，2022，35 (2)：231－239.

[3] 卫文韬，李亚军. 基于双流卷积神经网络的肌电信号手势识别方法 [J]. 计算机集成制造系统，2022，28 (1)：124－131.

[4] 田　勇，郭　莹，崔家栋，等. 基于近似熵子载波选择的人体手势识别方法 [J]. 计算机工程与设计，2022，43 (2)：323－329.

[5] 王　帅，张淑军，叶　康，等. 基于改进 Transformer 的连续手语识别方法 [J]. 计算机科学，2022，49 (S2)：573－578.

[6] 杨观赐，韩海峰，刘赛赛，等. 基于全局注意力机制和 LSTM 的连续手语识别算法 [J]. 包装工程，2022，43 (8)：28－34.

[7] 杨光义，丁星宇，高　毅，等. 基于注意力机制的复杂背景连续手语识别 [J]. 武汉大学学报（理学版），2023，69 (1)：97－105.

[8] MELEK A，LSHAK P，KENAN C. Real-time sign language recognition based on YOLO algorithm [J]. Neural computing & applications，2024，36 (14)：7609－7624.

[9] 安微微，武忆涵，王高峰. 一种改进的鱼眼畸变校正算法 [J]. 计算机测量与控制，2023，31 (10)：182－187.

[10] 兰颖华，王　鉴，韩　焱. 基于鱼眼相机畸变图像的大尺寸目标测量方法 [J]. 电子测量技术，2022，45 (19)：161－166.

[11] 郭乐铭，薛万利，袁甜甜. 多尺度视觉特征提取及跨模态对齐的连续手语识别 [J]. 计算机科学与探索，2024，18 (10)：2762－2769.

[12] KALIYAPERUMAL G V，GOPALAN A P. A deep neural network framework for dynamic Two-Handed indian sign language recognition in hearing and Speech-Impaired communities [J]. Sensors，2025，25 (12)：3652－3652.

[13] ACUAPAN H G，HERNÁNDEZ O K，CASTELÁN M，et al. Toward a recognition system for mexican sign language：arm movement detection [J]. Sensors，2025，25 (12)：3636－3636.

[14] KIM T，KIM B. Enhancing sign language recognition performance through Coverage-Based dynamic clip generation [J]. Applied Sciences，2025，15 (11)：6372－6372.

[15] 孙雅静，郑　爽，罗显志. 基于改进 YOLOv8n 的手语识别算法 [J]. 湖北大学学报（自然科学版），2025，47 (4)：539－550.

[16] YU M，GAN A，XUE C，et al. DSTEN-CSLR：dual spatial-temporal enhancement network for continuous sign language recognition [J]. Neural Computing and Applications，2025，37 (19)：1－24.

[17] 张守震，姜　飞，郭　都，等. 基于深度学习模型的手语识别算法研究及应用 [J]. 兰州文理学院学报（自然科学版），2025，39 (2)：64－67.

[18] LIANG S，LI Y，SHI Y，et al. Integrated multi-local and global dynamic perception structure for sign language recognition [J]. Pattern Analysis and Applications，2025，28 (2)：50－50.

[19] 闵越聪，陈熙霖. 面向连续手语识别的自适应关键帧选择 [J]. 中国科学：信息科学，2024，54 (4)：893－910.

[20] 陈凯喆，高煜新. 基于聋哑群体的手语图形设计研究 [J]. 明日风尚，2023，(24)：116－118.

[21] 黄同愿，谭　禹，朱金江. 基于 SlowFast 网络的孤立词手语识别算法研究 [J]. 重庆理工大学学报（自然科学），2023，37 (12)：267－275.