计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(4):375−380 [doi: 10.15888/j.cnki.csa.008397]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于二阶隐马尔可夫模型的连续手语识别①

梅家俊, 王卫民, 戴兴雨

(江苏科技大学 计算机学院, 镇江 212100)
通信作者: 梅家俊, E-mail: 895215236@qq.com

摘　要: 在传统的一阶隐马尔可夫模型 (HMM1) 中, 状态序列中的每一个状态被假设只与前一个状态有关, 这样虽然可以简单、有效地推导出模型的学习和识别算法, 但也丢失了许多从上文传递下来的信息. 因此, 在传统一阶隐马尔可夫模型的基础上, 为了解决手语识别困难、正确率低的问题, 提出了一种基于二阶隐马尔可夫模型 (HMM2) 的连续手语识别方法. 该方法利用滑动窗口算法使手语视频切分成多个手语短视频, 通过三维卷积模型得到手语短视频和手语词汇视频的特征向量, 由此计算出二阶隐马尔可夫模型的相关参数, 并运用 Viterbi 算法实现连续手语的识别. 实验证明, 基于二阶隐马尔可夫模型的手语识别取得了 88.6% 的识别准确率, 高于传统的一阶隐马尔可夫模型.

关键词: 手语识别; 滑动窗口; 二阶隐马尔可夫; 三维卷积; Viterbi; 深度学习; 卷积码

引用格式: 梅家俊,王卫民,戴兴雨.基于二阶隐马尔可夫模型的连续手语识别.计算机系统应用,2022,31(4):375–380. http://www.c-s-a.org.cn/1003-3254/8397.html

Continuous Sign Language Recognition Based on Second-order Hidden Markov Model

MEI Jia-Jun, WANG Wei-Min, DAI Xing-Yu

(School of Computer Science, Jiangsu University of Science and Technology, Zhenjiang 212100, China)

Abstract: In the traditional first-order hidden Markov model (HMM1), each state in the state sequence is assumed to be only related to the previous state. In this way, although the model learning and recognition algorithm can be simply and effectively deduced, a lot of information passed down from the above is lost. Therefore, in view of the traditional HMM1, a continuous sign language recognition method based on the second-order hidden Markov model (HMM2) is proposed to solve the problems of the difficulty and low accuracy of sign language recognition. In this method, a video of sign language is divided into several short videos by the sliding window algorithm, and the feature vectors of the short videos and word videos of sign language are obtained through the 3D convolution model. The relevant parameters of the HMM2 are thereby calculated, and continuous sign language recognition is achieved via the Viterbi algorithm. Experimental results show that the accuracy of sign language recognition based on the HMM2 is 88.6%, which is higher than that of the traditional HMM1.

Key words: sign language recognition; sliding window; second-order hidden Markov; 3D convolution; Viterbi; deep learning; convolution code

手语识别技术其实是利用计算机等智能设备, 将手语动作转换为能够与其他社会群体交流的信息[1]. 目前中国手语大约有 5 500 多个常用词汇[2]. 然而真正理解手语含义的人却数量极少, 大部分非聋哑人对手语基本上一无所知, 并且也很少有人愿意去花时间和精力去学习手语这项技能, 这也是聋哑人群体与其他社会群体之

① 收稿时间: 2021-06-02; 修改时间: 2021-07-07; 采用时间: 2021-07-13; csa 在线出版时间: 2022-03-22

间产生沟通障碍的原因之一. 因此, 研究手语识别技术不仅能使聋哑人更好地适应社会环境, 还可以促进人机交互的发展, 为用户提供更好的人机交互体验[3].

本文将隐马尔可夫模型和深度学习结合起来, 其中由于深度学习具有很好的迁移能力, 故而将深度学习用于手语词汇的特征提取以及候选词汇的选择, 而隐马尔可夫模型可以将现有的语言学的模型应用到手语的识别中, 并且二阶隐马尔可夫模型可以更好地运用经验知识来辅助识别[4].

1 相关工作

根据识别手语的方法上划分, 可以将手语识别分为四类, 分别为: 基于体感设备的手语识别、基于穿戴式设备的手语识别、基于传统机器学习的手语识别, 以及基于深度学习的手语识别.

基于体感设备的手语识别就是采用体感设备 Kinect 来进行手语的识别. 如 2019 年, 千承辉等[5] 借助 Kinect 得到人体的深度图像以及骨骼特征信息提取出手部特征, 实现动态的手语识别, 准确率可达 95%, 2020 年, 陈德宁等[6] 运用 Kinect 设备提取到人体的骨骼信息, 对手语进行分类和识别.

基于穿戴式设备的手语识别, 即通过运用数据手套以及位置跟踪器获取手势的实时变化信息, 如 Oz 等[7] 使用 CyberGlove 数据手套以及位置跟踪器对 300 个美国手语单词进行识别, 得到了 90% 的识别准确率.

基于机器学习的手语识别, 就是使用传统的机器学习算法, 如隐马尔可夫模型[8], 支持向量机等[9]. 2019 年, 蒋贤维等[10] 使用精度高斯支持向量机 (FGSVM) 对中国手语手指语图像样本进行实验, 最终分类达到 92.7% 的识别率, 2020 年, 包嘉欣等[11] 使用综合多要素的手语肤色分割与改进 VGG 网络的手语识别方法, 得到了对手语图像 97% 以上的平均识别率.

深度学习方法的出现, 使手语识别迎来了新的机遇. 例如 2018 年, 梁智杰等[12] 运用三维卷积神经网络, 提取出视频的短时特征, 并运用双向 LSTM 输入到残差网络中, 最终得到了很好的性能表现.

然而, 无论在国内还是国外, 都很难在市场上看到成熟的手语识别设备, 主要原因是基于体感设备的手语识别, 虽然识别准确率很高, 但体感设备 Kinect 价格昂贵且体积较大, 对于非室内的手语识别环境不是很方便; 而基于穿戴式设备的手语识别方式, 虽然数据手套准确率高, 但手套设备昂贵, 且不易操作, 不便于携带, 很难普及和推广; 基于机器学习的手语识别, 虽然不需要配套的设备, 但整体的识别率比使用相关设备的识别方法低; 而基于深度学习的手语识别, 在训练量较少的时候效果并不佳, 只有在有大量数据的情况下, 深度学习方法才可以得到一个较好的识别准确率.

2 基于 HMM2 的连续手语识别方法

手语识别是将一段完整的手语视频转换为中文语义的过程[13]. 手语识别主要是由语料库和手语词汇视频库的建设、滑动窗口的切分及手语视频识别等技术组成.

连续手语识别的系统运算流程如图 1 所示.

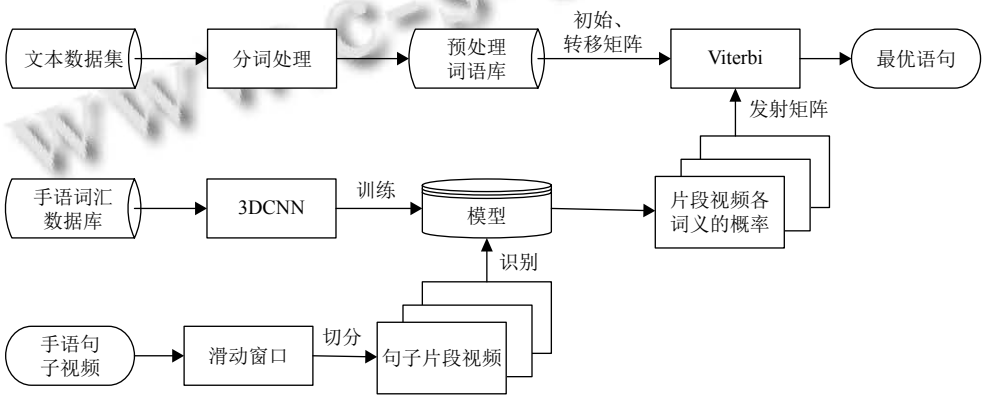

图 1 系统流程图

2.1 语料集和手语词汇视频库的建设

中文语料集以及手语词汇视频库的建设是手语识别中最基本也是最重要的环节. 通过收集大量的中文文本建立语料集, 再通过各种渠道, 大量收集手语词汇

的视频, 其中包括人、衣物、食品、生活用品、生活、工作、社会活动、哲学、伦理、心理、行为、事物的状态、性质、特点、民族、宗教、历史、政治、法律、国防外交、经济、文化教育、时间、空间、数学、物理、化学、信息学、生物等共 54 大类共 5 586 个手语词汇[14]. 把收集到的视频按各自表达的意思进行分类, 构建完整的手语词汇视频库[15].

2.2 滑动窗口的切分

由于手语语句是由多个手语词汇组合而成, 因此需要把手语视频进行切分, 分割成多个长度较短的视频, 使每个视频能够代表一个词汇[16]. 因此采用滑动窗口对手语视频进行切分, 根据窗口大小 *Size* 以及滑动步数 *Step*, 将原视频 V 切分成多个相互重叠的视频片段 $V_1, V_2, \cdots, V_m$, 其中 m 是切分后的视频个数.

2.3 手语视频识别 (基于 HMM2)

2.3.1 初始化 HMM2

初始化隐马尔可夫模型的 3 个参数矩阵: 初始状态矩阵, 转移状态矩阵和发射矩阵, 以及两个序列: 状态序列和观测序列. 状态序列 $s=W_1, W_2, \cdots, W_n$, 其中 n 表示手语词汇个数, W_i 表示第 i 个手语词汇, 观测序列 $V=V_1, V_2, \cdots, V_m$, 其中 V_i 表示切分出的第 i 个视频. 主要计算过程如下所示.

(1) 文章分句

输入 L 篇文章, 对 L 篇文章进行如下处理: 使用标点符号"? . !"进行分句处理, 设分句后的结果为 $S_1, S_2, S_3, \cdots, S_y$, 其中 y 为 L 篇文章的句子总个数, S_i 为 L 篇文章的第 i 个句子.

(2) 计算初始状态矩阵

计算所有字在句首出现的频次, 设字序列为 $C_1, C_2, \cdots, C_x$, 对应的频次序列为 $Count(C_1), Count(C_2), \cdots, Count(C_x)$, 其中 C_i 表示第 i 个字, $Count(C_i)$ 为字 C_i 出现的频次, 于是有:

$$P(C_i) = \log\left(\frac{Count(C_i)}{y}\right), 1 \leqslant i \leqslant x \tag{1}$$

由此得到了初始状态矩阵:

$$\pi = (P(C_1)P(C_2)\cdots P(C_x)) \tag{2}$$

(3) 句子分字

对所有的句子进行频次统计: 把 y 个句子中相应的 $C_1, C_2, \cdots, C_x$ 替换为 $C_1, C_2, \cdots, C_x$ 的下标 1, 2, ⋯, x. 去除其他非 1, 2, ⋯, x 的词汇, 再把下标 1, 2, ⋯, x 替换回相对应的字, 由此就得到了分字的结果.

(4) 计算一阶转移矩阵

首先统计出 C_i 的频次 $Count(C_i)$ 以及 C_iC_j 的频次 $Count(C_iC_j)$, 其中 $Count(C_iC_j)$ 表示为字 C_i 后出现 C_j 的概率. 再计算出每个字后出现所有字的频次总和, 记为 $O_1, O_2, \cdots, O_x$, 其中 O_i 表示 C_i 后出现每个 C_j 的频次总和.

$$P(C_j|C_i) = \log\left(\frac{Count(C_iC_j)}{Count(C_i)O_i}\right), 1 \leqslant i, j \leqslant x \tag{3}$$

其中, $P(C_j|C_i)$ 表示为 C_i 到 C_j 的转移概率, 得到了一阶转移状态矩阵:

$$A_1 = \begin{pmatrix} P(C_1|C_1) & \cdots & P(C_1|C_x) \\ \vdots & \ddots & \vdots \\ P(C_x|C_1) & \cdots & P(C_x|C_x) \end{pmatrix} \tag{4}$$

(5) 计算二阶转移矩阵

统计出 $C_iC_jC_k$ 的频次, 记为 $Count(C_iC_jC_k)$, 其中 $Count(C_iC_jC_k)$ 表示为 C_iC_j 后出现 C_k 的概率. 再计算每两个字后出现所有字的频次总和, 记为 $Z_{11}, Z_{12}, \cdots, Z_{xx}$, 其中 Z_{ij} 表示 C_iC_j 后出现每个 C_k 的频次总和. 于是有:

$$P(C_k|C_iC_j) = \log\left(\frac{Count(C_iC_jC_k)}{Count(C_iC_j)Z_{ij}}\right), 1 \leqslant i, j, k \leqslant x \tag{5}$$

其中, $P(C_k|C_iC_j)$ 表示 C_iC_j 到 C_k 的转移概率, 由此得到了二阶转移状态矩阵:

$$A_2 = \begin{pmatrix} P(C_1|C_1C_1) & \cdots & P(C_x|C_1C_1) \\ \vdots & \ddots & \vdots \\ P(C_1|C_xC_x) & \cdots & P(C_x|C_xC_x) \end{pmatrix} \tag{6}$$

(6) 三维卷积模型

每个手语词汇 $W_1, W_2, \cdots, W_n$ 都包含 k 个视频, 将这些视频分别记为 $d_{11}, d_{12}, \cdots, d_{nk}$, 其中 d_{ij} 表示为第 i 个手语词汇的第 j 个视频. 将所有的手语词汇视频进行训练, 最终得到一个三维卷积模型. 卷积结构如图 2 所示.

(7) 计算发射矩阵

将切分之后得到的 m 个视频 $V_1, V_2, \cdots, V_m$, 利用训练好的三维卷积模型进行手语词汇识别, 其中第 i 个切分视频的描述内容为第 j 个手语词汇的概率记为 $P(W_{ij})$. 对每个切分视频按照概率大小 $P(W_{ij})$ 进行排序, 保留排好序的前 5 个 $P(W_{ij})$, 其余 $P(W_{ij})$ 赋值为 0. 将每个手语词汇对 m 个切分的视频所得的值求和得到 $h_1, h_2, \cdots, h_n$. 于是有:

$$P(W_{ij}) = \log\left(\frac{P\left(W_{ij}\right)}{h_j}\right), 1 \leqslant i \leqslant m, 1 \leqslant j \leqslant n \quad (7)$$

对所有 $P(W_{ij})$ 中的非数值型的结果赋值为无穷小, 由此得到了发射矩阵:

$$B = \begin{pmatrix} P(W_{11}) & \cdots & P(W_{1n}) \\ \vdots & \ddots & \vdots \\ P(W_{m1}) & \cdots & P(W_{mn}) \end{pmatrix} \quad (8)$$

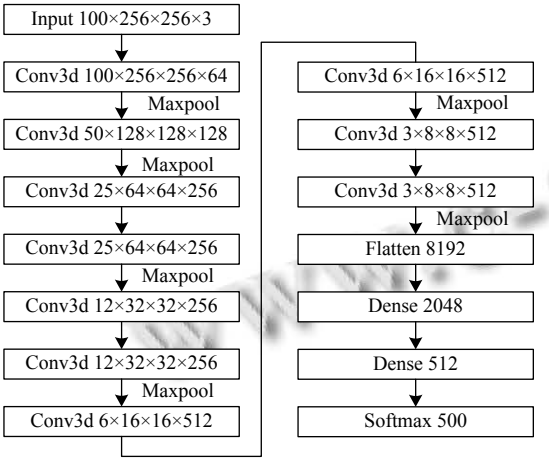

图 2 三维卷积网络结构

2.3.2 计算最佳路径

运用 Viterbi 算法计算 HMM2 的最佳路径, 主要过程如下所示:

给定一个观察序列 $O = V_1, V_2, \cdots, V_m$ 和 HMM2 模型 $\lambda=(\pi, A_1, A_2, B)$, 并选择一个状态序列 $S= W_1, W_2, \cdots, W_n$, 其中 V_i 表示第 i 个切分的视频, W_i 表示状态序列中的第 i 个词汇:

(1) 首先计算状态序列中的所有词汇作为句首的概率 $P(Wi)$, 从发射矩阵 B 中找出当观察值为 V_1 时, 相对应词汇的概率值 $B(Wi|O=V_1)$, 得到 Viterbi 算法中观测值为 V_1 的结果:

$$P(W_i|O = V_1 = P(W_i) \times B(W_i|O = V_1), i = 1,2,\cdots,n \quad (9)$$

(2) 接着计算当观测到 V_1 的状态值为 W_i 时, 观测到 V_2 的状态值为 W_j 的概率 $P(W_j|W_i)$, 从发射矩阵 B 中找出当观察值为 V_2 时, 相对应词汇的概率值 $B(W_j|O=V_2)$, 则得到 Viterbi 算法中观测值为 V_2 的结果:

$$P(W_iW_j|O = V_2) = P(W_i) * P(W_j|W_i) * B(W_j|O = V_2), \quad i, j = 1,2,\cdots,n \quad (10)$$

(3) 观测值为 V_t 时, 其中 $2 \leqslant t \leqslant m$, 计算出当 t–1 时刻观测值为 W_j 并且 t–2 时刻观测值为 W_i 时, 观测到 W_k 的概率 $P(W_k|W_iW_j)$, 从发射矩阵 B 中找出当观察值为 V_t 时, 相对应词汇的概率值 $B(W_j|O=V_t)$, 则得到 Viterbi 算法中观测值为 V_t 的结果:

$$P(W_iW_jW_k|O = V_t) = P(W_iW_j|O = V_{t-1}) * P(W_k|W_iW_j) * B(W_k|O = V_t), i, j, k = 1,2,\cdots,n \quad (11)$$

(4) 求取最佳序列状态: 找出 $P(W_iW_jW_k|O=V_m)$ 的最大值, 并记录下相对应的 W_i, W_j, W_k, 通过 W_i, W_j 依次找出前一个观测时刻的 W_i, 最终得到一条概率最大的路径, 即识别出的最终结果.

3 实验

3.1 实验数据来源

为了验证该系统的可行性, 本次实验从新浪、搜狐上爬取 50 篇新闻文章构成语料集, 并获取中国科学技术大学视觉手语研究小组 (VSLRG) 发布的中国孤立手语词数据集作为手语视频库, 该数据集由 50 个唯一的参与者执行, 每个参与者对每个类别进行 5 次. 数据集包含 500 个不同的类别, 例如: 身体、头部、头发、女士、妹妹、杯子、灯光等. 并且每个类别具有 250 个实例, 共计 125 000 个手语词汇视频, 对于 125 000 个手语词汇视频, 按照 9 比 1 的比例, 随机划分卷积模型的训练集和测试集, 训练集共 112 500 个词汇视频, 测试集共 12 500 个词汇视频. 再获取 VSLRG 发布的中国连续手语数据集, 该数据集包括 100 种不同含义的句子, 从每种句子中随机挑选 5 个视频, 共 500 个连续手语视频作为手语识别的测试集. 其中图 3 为连续手语数据集中的一个视频的关键帧, 图 4 为孤立手语词数据集的一个词汇视频的关键帧.

3.2 模型训练及手语识别

首先给定一个前提: 测试集中的每一个连续手语视频中, 表示手语的速度与手语词汇视频中, 某个人表示手语的速度是相似的.

每个手语词汇 $W_1, W_2, \cdots, W_n$ 都包含 k 个视频, 将这些视频分别记为 $d_{11}, d_{12}, \cdots, d_{nk}$, 其中 d_{ij} 表示第 i 个手语词汇的第 j 个视频, 并且每一个手语词汇视频都有不同长度的帧数, 记为 $F_{11}, F_{12}, \cdots, F_{nk}$, 其中 F_{ij} 表示为第 i 个手语词汇的第 j 个视频的帧数, 把所有视频的帧数统计出来, 如表 1 所示.

图 3 “他的同学是警察”的关键帧

图 4 手语词汇“外祖父”的关键帧

表 1 词汇视频帧数统计

词汇视频个数	最长词汇视频帧数	最短词汇视频帧数
125 000	239	22

由于最长词汇视频帧数过大, 测试集中部分连续手语视频达不到此帧数, 因此统计 100 帧以下的视频占比, 如表 2 所示.

表 2 100 帧以下词汇视频总数占比

词汇视频个数	100帧以下的视频个数	占比
125 000	121 893	97.51%

共有 97.51% 的视频处于一百帧以内, 并且对于连续视频而言, 切分的窗口过大, 反而会产生不必要的误差. 于是, 设置计算过程中词汇视频帧数为 100, 大于 100 帧的词汇视频取前 100 帧计算, 小于 100 帧的词汇视频补零补到 100 帧. 将处理好的训练集进行训练, 共计训练 1 000 轮.

为验证算法的识别性能, 将通过与其他算法, 如余弦相似度[17] 和二维卷积模型[18], 进行对比 Top-1 及 Top-5 的准确率.

其中网络模型如图 1 所示, 训练中设置批量大小 batch_size 为 64, 使用 Adam 优化器, 根据训练的效果, 依次调整初始学习率为 1×10^{-4}, 1×10^{-6}, 1×10^{-8}, 其他参数保持不变, 并设置训练迭代次数 epochs 为 1 000. 实验结果如表 3 所示.

表 3 不同算法的识别效果 (%)

算法	Top-1	Top-5
余弦相似度	32.8	54.6
二维卷积模型	70.9	89.4
本文卷积模型	76.1	94.7

对测试集中每一个连续手语视频, 进行如下处理:

以当前连续手语视频的前 100 帧作为第一个手语短视频, 利用训练好的三维卷积模型进行手语词汇识别, 找出其中概率最大的五个手语词汇, 按照这 5 个手语词汇的概率比例, 与每个手语词汇中 250 个词汇视频帧数的平均值进行计算, 得到滑动的帧数并进行滑动, 得到第二个手语短视频. 第二个短视频也进行词汇识别, 同样以概率最大的 5 个手语词汇计算出第二次滑动的帧数. 直至滑动得到的手语短视频的帧数小于 100 帧后滑动结束.

由此得到了当前连续手语视频的发射矩阵, 并根据语料集得到初始状态矩阵, 一阶转移矩阵以及二阶转移矩阵. 最终通过 Viterbi 算法进行识别, 结果如表 4 所示.

表 5 为本文算法 HMM2 与传统一阶隐马尔可夫 HMM1 的对比情况.

表 4 识别效果

真实文本	预测文本
他的同学是警察	他的同学是警察
他外祖父是邮递员	他外祖父是民政

表 5 手语视频识别精度

算法	正确识别的视频个数	识别精度 (%)
HMM1	409	81.8
HMM2	443	88.6

4 结语

本文采用了二阶隐马尔可夫模型与深度学习相结合的方式, 其中深度学习用于词汇的特征提取以及候选词汇的选择, 极大地缩小了候选词汇的范围, 而二阶隐马尔可夫能够更好地运用经验知识来辅助识别, 通过结合深度学习得到的候选词汇, 形成连续手语视频的句义. 在连续手语数据集上实验结果表明, 本文算法的 Top-5 识别精度为 94.7%, 比二维卷积模型提高 5.3%; 二阶隐马尔可夫模型对于连续语句的识别精度为 88.6%, 比一阶隐马尔可夫提高 6.8%, 具有更好的识别精度.

参考文献

1 冯欣. 基于 Kinect 的非特定人连续中国手语识别 [硕士学位论文]. 青岛: 山东大学, 2018.

2 毛晨思. 基于卷积网络和长短时记忆网络的中国手语词识别方法研究 [硕士学位论文]. 合肥: 中国科学技术大学, 2018.

3 郭鑫鹏, 黄元元, 胡作进. 基于关键帧的连续手语语句识别算法研究. 计算机科学, 2017, 44(S2): 178–183. [doi: 10.11896/j.issn.1002-137X.2017.11A.037]

4 Sung-Hyun Y, Thapa K, Kabir MH, *et al*. Log-Viterbi algorithm applied on second-order hidden Markov model for human activity recognition. International Journal of Distributed Sensor Networks, 2018, 14(4): 155014771877254.

5 千承辉, 邵晶雅, 夏涛, 等. 基于 Kinect 的手语识别方法. 传感器与微系统, 2019, 38(6): 31–34, 38. [doi: 10.13873/J.1000-9787(2019)06-0031-04]

6 陈德宁, 马锐军, 张俊源, 等. 基于 Kinect 的手语识别及播放器设计. 科技风, 2020, (14): 23–24. [doi: 10.19392/j.cnki.1671-7341.202014019]

7 Oz C, Leu MC. American Sign Language word recognition with a sensory glove using artificial neural networks. Engineering Applications of Artificial Intelligence, 2011, 24(7): 1204–1213.

8 丰月姣, 贺兴时. 二阶隐马尔科夫模型的原理与实现. 价值工程, 2009, 28(12): 103–105. [doi: 10.3969/j.issn.1006-4311.2009.12.033]

9 曹芳宁. 基于 Zernike 矩和支持向量机的手势识别研究 [硕士学位论文]. 南京: 南京大学, 2017.

10 蒋贤维, 张妙娴, 朱兆松. 基于灰度共生矩阵和精度高斯支持向量机的中国手语手指语识别. 计算机科学, 2019, 46(S2): 303–308.

11 包嘉欣, 田秋红, 杨慧敏, 等. 基于肤色分割与改进 VGG 网络的手语识别. 计算机系统应用, 2020, 29(6): 47–55. [doi: 10.15888/j.cnki.csa.007448]

12 梁智杰, 廖盛斌. 融合宽残差和长短时记忆网络的动态手势识别研究. 计算机应用研究, 2019, 36(12): 3846–3852. [doi: 10.19734/j.issn.1001-3695.2018.07.0429]

13 李姝蓉. 基于视频的动态手语识别算法研究 [硕士学位论文]. 南京: 南京航空航天大学, 2014.

14 金力. 基于移动互联网的手语翻译器的设计与实现 [硕士学位论文]. 镇江: 江苏科技大学, 2017.

15 张轶秋, 王卫民, 唐洋, 等. 基于状态机的手语动画自动生成技术. 计算机与数字工程, 2020, 48(1): 217–220. [doi: 10.3969/j.issn.1672-9722.2020.01.041]

16 Ji Y, Zhong J. Improved HOG feature vehicle recognition algorithm based on sliding window. Journal of Physics: Conference Series, 2020, 1627(1): 012013.

17 艾佳琪, 左毅, 刘君霞, 等. 基于余弦相似度的动态语音特征提取算法. 计算机应用研究, 2020, 37(S2): 147–149.

18 李晨, 黄元元, 胡作进. 基于深度学习的连续手语语句识别算法. 计算机技术与发展, 2021, 31(1): 1–6. [doi: 10.3969/j.issn.1673-629X.2021.01.001]