华东理工大学学报(自然科学版)
Journal of East China University of Science and Technology
ISSN 1006-3080,CN 31-1691/TQ

《华东理工大学学报(自然科学版)》网络首发论文

题目：面向手语识别的视频关键帧提取和优化算法
作者：周舟，韩芳，王直杰
DOI：10.14135/j.cnki.1006-3080.20191201002
收稿日期：2019-12-01
网络首发日期：2020-09-23
引用格式：周舟，韩芳，王直杰．面向手语识别的视频关键帧提取和优化算法[J/OL]．华东理工大学学报(自然科学版).
https://doi.org/10.14135/j.cnki.1006-3080.20191201002

DOI: 10.14135/j.cnki.1006-3080.20191201002

面向手语识别的视频关键帧提取和优化算法

周　舟，　韩　芳，　王直杰
（东华大学信息科学与技术学院，上海 201620）

摘要：基于计算机视觉的手语识别技术可以为聋校的双语教学带来很大的便利，而手语识别技术的难点之一在于视频关键帧的提取。根据手语视频关键帧的特点和手语者的手语习惯，提出了一种面向手语识别的视频关键帧提取和优化算法。首先利用卷积自编码器提取视频帧的深度特征，对其进行K-means聚类，在每类视频帧中采用清晰度筛选取出最清晰的视频帧作为初次提取的关键帧；然后利用点密度方法对初次提取的关键帧进行二次优化，得到最终提取的关键帧进行手语识别。实验结果表明，本文算法能大量消除冗余帧，并能提高手语识别的准确率和效率。

关键词：手语识别；视频关键帧；卷积自编码器；K-means；点密度
中图分类号：TP391　　**文献标志码：**A

在聋校语言教学中，双语教学模式可以有效地提高聋童的语言学习效率，但对特殊教师来说则需花费更多的耐心、时间和精力。目前，我国特殊教育学校的师资普遍面临着薄弱的现状，利用手语识别技术可帮助这些教师完成教学任务。即聋童将手语录成视频输入计算机，进而能够学习输出的汉字和唇语，无需老师亲自教学就能完成汉语书面语的学习。另外，计算机是针对标准的手语（以《中国手语》为标准）进行识别，借此还可纠正聋童手语方言化的问题。

基于视觉的手语识别技术是一个富有挑战性的、多学科交叉的研究课题，是人机交互领域的一个前沿性课题和研究热点。一个手语动作的视频中并不是每一帧都对其表达的语义有作用，由于手语者潜意识里会强调手语语义，在做手语动作时手部会在关键手势处有短暂的停顿，此处包含该手语的重要语义，这类包含关键手势的视频帧称为关键帧。如何从一段手语视频中检索出有效的关键帧直接影响手语识别算法的性能。

视频关键帧提取方法一般分为四大类：第1类是基于图像内容的方法[1]。该方法将视频内容变化程度作为选择关键帧的标准，而其中视频内容主要由图像的特征体现。文献[2]中对图像底层特征（如颜色直方图、颜色矩、惯性矩、边缘等）进行加权融合，用于筛选关键帧。第2类是基于运动分析的方法[3]-[4]。一般思想是计算出每帧图像的光流场，然后对光流图进行计算，极小值对应的那一帧被选为关键帧[5]。该方法能很好地表达视频内的全局性运动，但计算量较大，耗时较长。这两类方法都没有使用更具特征表达能力的深度图像特征，所以效果不佳。考虑到动态手语的特点和手语者的心理意识，部分研究者[6]提出了第3类基于轨迹曲线点密度特征的关键帧检测算法。利用轨迹密度曲线上点的密度大小区分关键帧与非关键帧，但有时会由于手心定位不准而产生轨迹偏差大，对关键帧的提取影响较大。第4类是目前的主流方法——基于聚类的方法[7]-[8]。该方法预先设定好聚类数目，将相似的帧聚为一类，每一类代表一个关键帧，但此类方法提取的关键帧往往存在大量的冗余。文献[9]在该方法的基础上根据帧间位置对初始关键帧序列进行二次优

收稿日期：2019-12-01
基金项目：国家自然科学基金（11572084, 11972115）；中央高校基本科研业务费专项资金；东华大学"励志计划"（18D210402）
作者简介：周　舟（1996—），女，湖北黄冈人，硕士生，主要研究方向为机器视觉、目标检测、图像处理。E-mail: 747346239@qq.com
通信联系人：韩　芳，E-mail: yadiahan@dhu.edu.cn

化,消减冗余信息构建最优关键帧序列,但对于画面每秒传输帧数(FPS)低的视频,上述方法可能会提取出手部较为模糊的关键帧,这会影响下一步的手势定位和识别。为此,本文在K-means聚类提取关键帧的基础上,使用清晰度筛选并结合点密度方法进行二次优化,以提高关键帧质量,消除冗余信息构建最优关键帧序列,并在中科院SLR Dataset中的xf500_color_video数据集进行了实验。

1 关键帧提取和优化算法

关键帧提取和优化算法的流程如图1所示,分为特征提取、K-means聚类、清晰度优化、点密度优化4个关键部分。

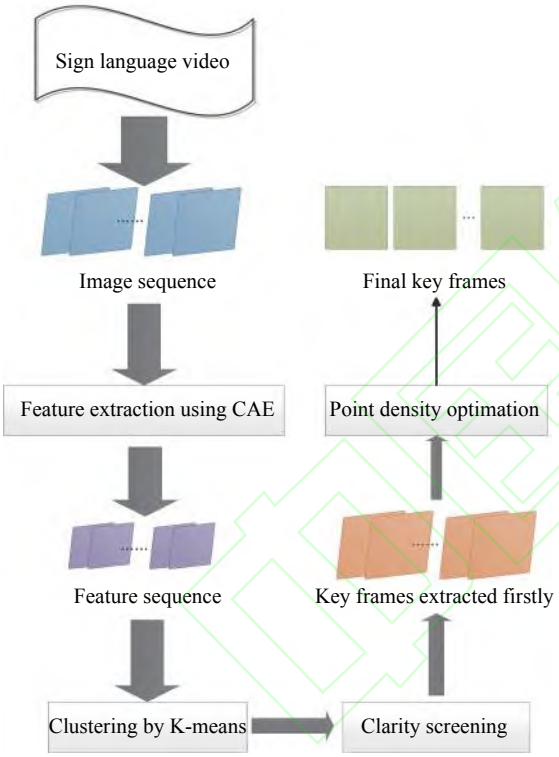


图1 关键帧提取算法流程图

Fig. 1 Flow chart of key frame extraction algorithm

1.1 特征提取

传统的K-means聚类提取视频关键帧的方法是计算各帧图像在像素空间的欧氏距离来衡量图像之间的差异,从而完成图像聚类。这种传统方法运用的是图像像素级特征,然而单个像素不能携带足够的图像语义信息,容易受到噪声的影响,而且会因图像尺寸大而导致计算量大。因此,本文考虑在特征空间中计算欧氏距离来衡量图像之间的差异,即采用自编码器网络[10]的一种——卷积自编码器(Convolutional Auto-Encoders, CAE)[11]来实现手语视频每一帧的深度特征提取。自动编码器主要应用于图像重构[12]、图像压缩[13]和特征提取等,通过编码器将输入数据压缩到一个潜在表示空间里,然后再通过解码器将潜在空间的数据进行重构,从而得到和输入相似的输出数据。CAE网络结构如图2所示。采用卷积层代替简单自编码器中的全连接层,利用卷积神经网络的卷积和池化操作,对输入的图像进行下采样以提供较小维度潜在表示,实现无监督特征提取。

本文中CAE的训练集是从SLR Dataset中的彩色视频中随机截取的视频帧。考虑到实验操作的便利,将所有的训练数据统一变为252 × 252的灰度图像,归一化处理后作为训练集,在无监督训练下训练。网络训练好后,对尺寸大小为1 080 × 720的手语视频的每一帧作同样的数据预处理,输入网络中的编码器后可得到56 × 56的能代表该帧重要特征属性的二维特征向量,作为K-means聚类的输入。

1.2 清晰度优化的K-means聚类

K-means聚类算法是聚类分析中运用最为广泛的算法之一。无论在数据处理或是图像、视频处理中,都运用得相当广泛,其基本思想是将相似对象归到同一簇,将不相似的对象归到不同簇。本文利用该算法对提取出的视频帧的深度特征进行聚类,将每一帧的图像特征展开成一个56 × 56 = 3 136维的向量,通过聚类可以得到K簇3 136维的特征向量。假如视频的特征序列为$\boldsymbol{F}=\{\boldsymbol{x}_0,\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$,$\boldsymbol{x}_i\in\mathbf{R}^N$,其中$n$为视频序列总帧数;$i$为视频中的第$i$帧,$\boldsymbol{x}_i$为视频中第$i$帧的特征向量,$\boldsymbol{x}_n$是$m$维向量,其中$m=3\,136$。对于绝大多数手语,一个关键手势的帧数不会超过6帧,那么本文选取聚类数目$K=\mathrm{len}(\boldsymbol{F})/6$,其中$\mathrm{len}(\boldsymbol{F})$为$\boldsymbol{F}$中特征向量的个数。具体算法如下:

(1)从$\boldsymbol{F}$中随机选取K个聚类质心(Cluster centroids),记为$\boldsymbol{u}_1,\boldsymbol{u}_2,\ldots,\boldsymbol{u}_k,\boldsymbol{u}_j\in\mathbf{R}^n(0<j<k)$,其中$u_j$为第$j$类的聚类质心。

(2)依次计算各个点到每个聚类质心的欧氏距离。定义样本$\boldsymbol{x}_i$到质心$\boldsymbol{u}_j$的欧氏距离为$\boldsymbol{D}_{ij}=\left\|\boldsymbol{x}_i-\boldsymbol{u}_j\right\|$,记集合$\boldsymbol{D}_i=\{D_{i1},D_{i2},\ldots,D_{ik}\}$,选取$\boldsymbol{D}_i$中的最小值$\boldsymbol{D}_{ij}$,此时将$\boldsymbol{x}_i$归入第$j$类。

(3)再对第j类的所有样本取均值,重新计算该类的质心。

(4)重复步骤(2)、(3),直到上个质心与重新计算的质心的差距最小。

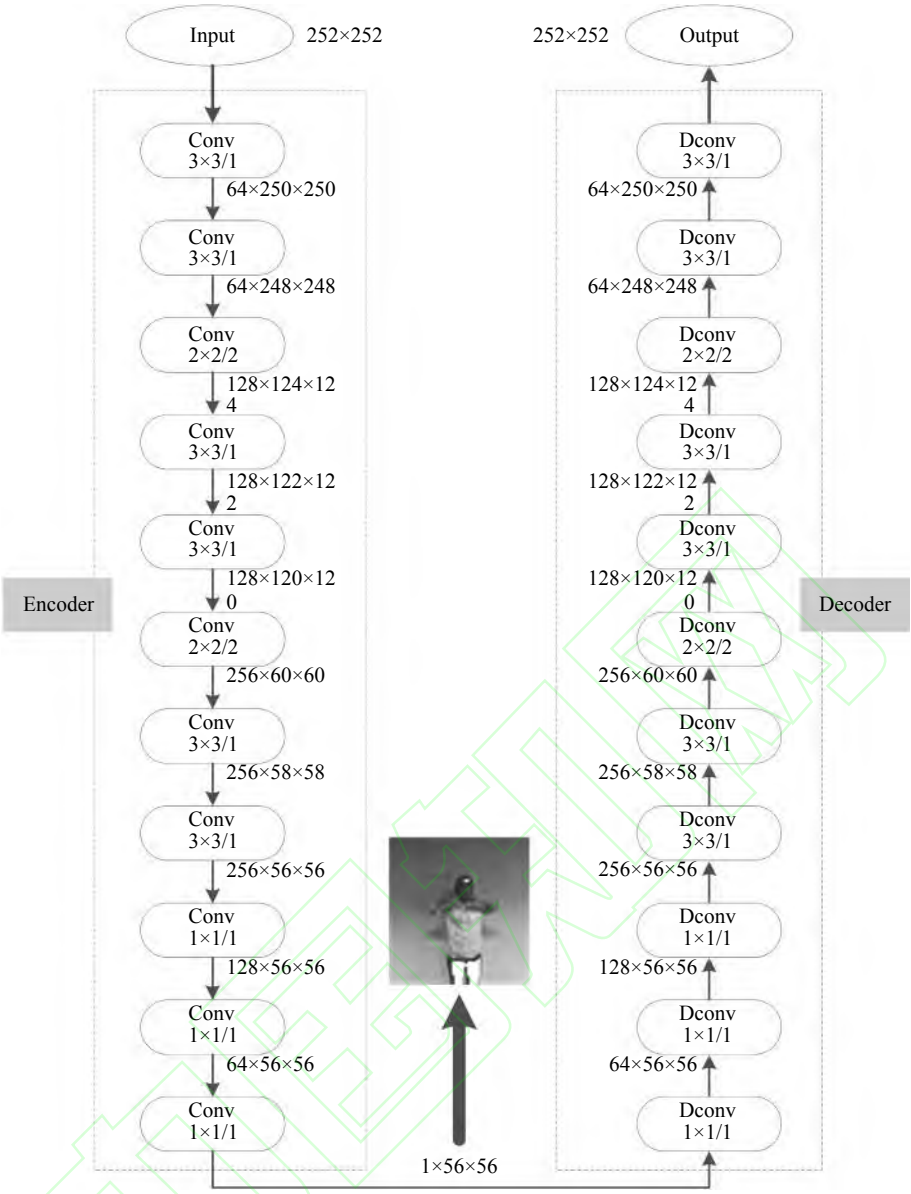


图 2　用于特征提取的 CAE 网络结构

Fig. 2　CAE network structure for feature extraction

手语“丑”的关键手势在《中国手语》中的定义如图 3 所示。该手语中手势可能在同一地点不同时间出现, 时间上相隔较远的两帧可能会出现在同一类中, 例如图 4 中的第 25 帧和 40 帧。一般来说, 相似的图像具有连续性, 则每一类中会包含连续帧段。考虑到有些手语中会有相同但不相邻的关键手势, 如果将其盲目归为一类, 则有可能造成手语语义的缺失。将聚得的每类中的非连续帧段分开另为一类, 再由时间先后将所有类排序。例如第 2 类应被分为 [18, 19, 20, 21, 22, 23, 24, 25] 和 [38, 39, 40] 两类, 此时既可以得到完整的相似帧的分类结果, 又没有丢失运动序列的重要信息。

根据 K-means 聚得的每类中包含的帧为相似的帧, 那么应选取其中最清晰的一帧作为最能代表该类的关键帧。文献 [14] 提出图像的清晰度可以由其边缘来判断, 图像的灰度级突变越大, 图像边缘特征越明显, 图像越清晰, 其目标物体与背景边缘越锐

图 3　手语“丑”的关键手势定义

Fig. 3　Key gesture definition of sign language "ugly"

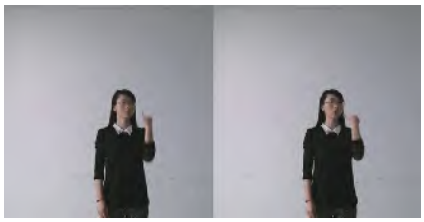


图4 手语“丑”中处于同一类中非相邻的两帧——(24帧和40帧)

Fig. 4 Two non-adjacent frames in the same class of sign language "ugly"

利,反之图像越模糊。图像的梯度可以很好地反映图像中目标物体边缘灰度,基于Tenengrad梯度函数的图像清晰度定义如下:

$$G(x,y)=\sqrt{G_x^2(x,y)+G_y^2(x,y)} \tag{1}$$

其中:$G_x(x,y)$和$G_y(x,y)$分别是图像像素点(x,y)处的像素灰度值$f(x,y)$与Sobel水平和垂直方向边缘检测算子的卷积,本文中使用的Sobel梯度算子为

$$g_x=\begin{bmatrix}-1&0&1\\-2&0&2\\-1&0&1\end{bmatrix},\ g_y=\begin{bmatrix}1&2&1\\0&0&0\\-1&-2&-1\end{bmatrix} \tag{2}$$

取每一类中$G(x,y)$值最大的一帧便得到首次提取的关键帧序列。对于大多数手语词语,包含同一关键手势的帧数不超过6帧,也就是说,这6帧里一般不会包含两个不同手势的关键帧,因此应在提取出的关键帧上再进行帧间隔优化,即相邻帧如果相距小于一个阈值a(本文中取$a=3$),则去掉这相邻两帧中序列号较小的一帧,这样有效地去除了首次提取关键帧的小部分冗余过渡帧。

1.3 点密度二次优化

本文定义关键手势为《中国手语》中每个词语对应的手势,那么手语视频中能清晰地显示这些手势的视频帧即为关键帧。本文的目的是希望最终提取的众多关键帧中存在包含这个手势的视频帧,并且尽可能减少动作过渡帧和不清晰帧的数量。初次提取的关键帧有大量的重复帧,对比这些重复的关键帧可以发现,一般是由于动作过于缓慢,最终导致本应归为一类的相似的两帧相似度变小。根据大量的统计实验发现,手语者做出关键动作时,手部往往会有下意识的短暂停留,导致K-means方法提取出的关键帧在关键手势附近比较密集。因此本文利用点密度对提取出的关键帧进行二次提取,优化关键帧序列,具体方法如下:

(1)依次记录1.2节中提取出的关键帧数量M,将这些关键帧在视频中的位置序列号记为p,并得到视频关键帧位置序号数组$p=\{p_i|i=1,2,\cdots,M\}$;

(2)依次计算p中每个点的点密度。定义第j个点p_j的点密度:

$$\text{Density}(p_j)=\{p_i|\text{dis}(p_j,p_i)<\delta,p_i\in p\} \tag{3}$$

其中,$\text{dis}(p_j,p_i)$表示p_j和p_i之间的欧氏距离。式(3)用于计算在p中有多少个点与p_j之间的距离小于阈值δ,满足条件的p_i越多,表示p_j的点密度越大,反之则表示p_j的点密度越小。定义阈值δ为p上所有相邻点之间的距离之和的平均值,即

$$\delta=\frac{\sum_{i=1}^{M-1}\text{dis}(p_i,p_{i+1})}{M-1} \tag{4}$$

以手语“丑”为例,从手语视频“丑”中按照1.2节的方法提取出的关键帧一共有7帧,即$M=7$,计算得到的点密度散点图如图5(a)所示。理论上,点密度大于阈值的点即为符合要求的关键帧,因此我们还需设定一个阈值T。阈值的选取可以有以下几种方式:(1)点密度的中位数;(2)所有点密度的平均值;(3)点密度最大值与最小值的均值。经过多次实验,本文选取所有点密度的平均值作为阈值T,如图5中$T=2.14$。超过阈值的点并非全部对应关键帧,并且对于大多数手语词语,包含关键手势的帧数不超过5帧,而有相邻两帧超过阈值的点的水平距离小于5帧的情况不少,此处必定会产生冗余的过渡帧或者重复的关键手势帧,因此对点密度图按照5帧的间隔进行等间隔划分,在每个区间中认定大于阈值T的第一个最大值对应的点为关键帧。如图5(b)所示,被划分的区间应为1.2节中进行K-means聚类的帧的序号区间,即6~68帧,最后获得的关键帧为空心圆圈出的点,一共2帧。

从图5中可以看出关键帧的数量减少了很多,但由于动作的连贯性,仍有重复的关键帧出现,但并未漏帧,只要保证未漏掉关键帧就可保证语义的完整性。图5中完整包含了手语词汇“丑”的所有关键帧。

2 关键手势识别

为了识别手语的含义,还需对得到的关键帧进行手势识别,本文采用SSD(Single Shot MultiBox)目标检测网络对手势进行定位和分类。SSD是近年来在目标检测中识别率较高、速度较快的算法,并且在小物体检测上尤为出色,其网络结构如图6所示。它使用VGG16作为基础网络,本文中该网络的输入为300 × 300的图像,分别在基础网络中的6个特征层

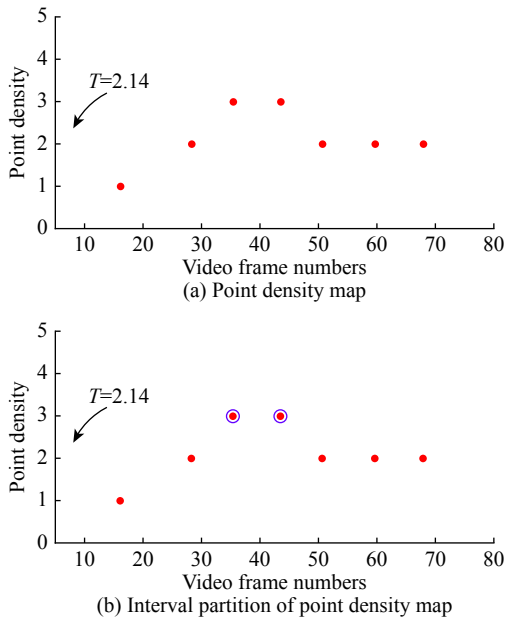


图 5　手语“丑”的点密度图

Fig. 5　Point density map of sign language "ugly"

上预测目标, 每个特征层都负责预测特定尺寸的目标, 整个网络通过端对端的训练方案在所有预测图上用定位损失和分类损失的加权进行优化。SSD 在 6 个特征层上将整幅图像分割成网格单元, 在每个网格单元中围绕中心点构造不同尺寸的边界框用于检测物体, 再将不同特征层中所有边界框结合起来, 经过 NMS(非最大值抑制)方法抑制一部分重叠或者不正确的边界框, 最终生成的边界框集合即为检测结果。

SSD 的训练集由多个手语视频的关键帧组成, 在训练 SSD 时, 把每种手势都作为一个输入类别, 那么输入的总类别有 C(手势类别数)+1(背景)个, 这样就可以通过 SSD 检测出每一帧图像中的每只手的手势类别, 再通过手势类别序列与由关键手势组成的标准模板库进行匹配, 得到该序列所表示的手语汉字。

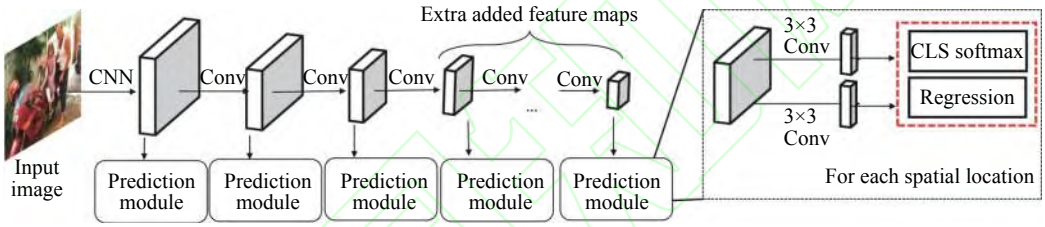


图 6　SSD 网络结构图

Fig. 6　Network structure of SSD

3　实验分析

实验选用中科院 SLR Dataset 数据集中 10 种不同手语词汇的彩色视频, 每个手语词汇由 50 个不同的手语者录制, 每个手语时长都不相同, 平均时长为 3~5 s, 拍摄条件相同, 所有视频的采样频率都为每秒 30 帧, 实验平台为 Python3.6。

为了验证 CAE 深度特征提取、清晰度优化和点密度优化在本文算法中的作用, 实验选取手语“松”的 50 个视频作为测试集, 平均每个视频的帧数为 37.42 帧, 关键帧有 2 个。在测试集上进行消融实验的结果见表 1。其中方法 1、2、3 分别表示在本文算法的基础上单独去掉 CAE 特征提取、清晰度优化、点密度优化的方法。

表 1　消融实验结果对比

Table 1　Comparison of ablation experimental results

Method	CAE feature extraction	Clarity screening	Point density	Numbers of key frames	Average detection time/s	Average detection accuracy/%
1		√	√	2.80	6.09	87.6
2	√		√	2.71	5.78	86.8
3	√	√		4.68	5.63	87.3
This paper	√	√	√	2.54	5.64	88.1

从表 1 中可以看出, 每个部分都能在识别效果上单独起到作用。其中方法 1 直接使用 720×720 的图像进行聚类, 而本文方法则是先进行 CAE 特征提取, 再利用 65×65 的特征向量进行聚类, 由于聚类的计算复杂度会随着输入样本的尺寸而改变, CAE 特征提取也会消耗时间, 因此在识别时间上不会呈现数量级的差异。方法 2 相比于本文算法提取到的关键帧不太清晰, 因此准确率受到相对大的影响。方

法3省略了点密度二次优化,所提取到的关键帧数量相比本文算法明显增多,并会因冗余帧的误识别影响识别准确率。

对同样的样本进行关键帧提取,并利用SSD网络分别对3种方法得到的关键帧进行手势识别。本组实验利用数据集中的500个手语视频进行测试,对每个手语单词的结果取平均值,结果见表2。从表2中可以清晰地观测到,与另外两种方法相比,本文算法提取的关键帧数量大大减少,因此在对关键帧手势识别上消耗的时间更少,所以识别时间远优于其他方法;其次,在另外两种方法上提取的关键帧包含多张冗余帧,即包含不清晰关键帧和非关键帧,因此在手势识别过程中会产生错误分类的影响,导致手势识别失败,识别准确率下降。

表2 3种算法对于不同手语单词的实验结果

Table 2 Experimental results of three algorithms for different sign language words

Words	Total frame numbers	This paper				Literature [9]			Literature[15]		
		numbers of keyframes	Numbers of keyframes	detectionti me/s	accur-acy/%	Numbers of keyframes	Detectionti me/s	Accur-acy/%	Numbers of keyframes	Detectionti me/s	Accur-acy/%
ugly	44.4	1	2.24	4.23	90.6	5.54	4.91	89.5	6.12	5.39	89.4
loose	37.4	2	2.54	5.64	88.1	4.42	6.54	87.0	4.52	7.19	86.9
ban	37.7	1	1.72	3.62	90.1	4.32	4.20	88.6	4.54	4.62	88.9
aunt	44.2	2	2.21	4.26	89.3	5.41	4.94	88.2	5.64	5.43	88.1
sky	48.2	1	2.54	4.34	91.0	5.94	5.03	89.5	6.64	5.53	89.8
tailor	75.4	2	3.66	5.01	86.2	10.2	5.81	85.6	10.7	6.39	85.0
status	50.0	2	2.52	4.23	87.6	6.23	4.91	86.5	6.64	5.39	86.4
baozi	39.5	2	1.82	3.56	84.3	4.54	4.13	83.4	4.84	4.54	83.1
we	52.7	2	2.48	4.52	89.6	6.84	5.24	88.5	7.16	5.76	88.4
propi-tious	70.9	2	3.96	4.32	89.4	9.86	5.01	88.3	10.8	5.51	89.2

目前,对于手语视频关键帧的提取效果还没有统一的指标进行判定,本文采用压缩率、误检率和漏检率3个参数来检验视频关键帧的提取效果,定义如下:

压缩率=提取关键帧数/总帧数

误检率=不包含关键手势的关键帧/总关键帧数

漏检率=没检测到的关键手势数/定义的关键手势数

压缩率越大说明该算法消除冗余帧越多,提取的关键帧更具代表性。提取出的视频关键帧将作为后期手势识别模块的输入,并且在提取出的视频关键帧中不包含关键手势的视频帧占比越大,在手势识别时消耗在无效关键帧的时间越多,有时还会因为被识别成错误的手势导致手语识别过程的成功率下降。因此误检率越小,说明关键帧提取效果越好。提取出的关键帧中缺失任意一个重要手势都会导致手语识别失败,所以为了提高手势识别的成功率,漏检率越小越好,表明算法的稳定性也越高。因此在实验时只需要检测以上3个指标即可充分衡量算法的优劣。对数据集中500个视频的3个指标求取平均值,实验结果如表3所示。

表3 3种算法的实验结果

Table 3 Experimental results of three algorithms

	Compression ratio/%	False detection rate/%	Missed detection rate/%
The porper	4.9	44.17	9.86
Literature [9]	12.5	69.97	71.31
Traditional K-means[15]	13.3	71.31	11.08

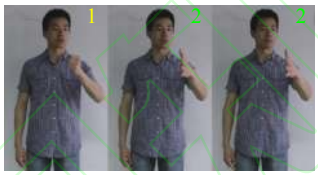
从实验结果可以看出,本文算法的压缩率和误检率相对另外两种算法有很大优化,一方面极大地减少了冗余关键帧,包含关键手势的帧占总结果帧数的比例高;另一方面,提取出的正确关键帧的能力更强,漏检率相比其他方法没有增大,因此该算法也保持了良好的稳定性。

图7示出了利用3种算法从一位手语者做手语“松”动作时提取的关键帧。其关键手势有2个,即一手握拳,之后再一手完全张开,图中的“1”表示该帧属于第一帧关键帧,“2”则表示该帧属于第二帧关键

帧。可以看出图 7(a)中由文献 [15] 提取出的关键帧最多, 相比之下, 图 7(b)中文献 [9] 对提取的关键帧二次优化后, 关键帧数量减少了 2 帧, 虽然减少了一定的冗余帧, 但一般由 K-means 聚类得出的关键帧较为分散, 相邻小于 3 帧的情况较少, 因此文献 [9] 的优化算法在冗余关键帧去除效果上并不明显。而本文针对一般人在做手语关键动作时的下意识的停顿动作, 导致 K-means 在停顿时间较长处提取出的关键帧的点密集大这一问题, 利用点密度筛选出停顿时间较长处即点密度大的视频帧, 更有目的性地提取关键帧, 因此除去的冗余帧更多。由图 7(c)可看, 出本文算法提取出的视频帧只有 3 帧, 其中完整地包括了两个关键手势。第二个关键手势有重复帧, 但在后期识别时, 一旦一个手语单词视频中识别出相同的手势则自动略去, 不会影响识别结果, 因此本文的结果在 3 种方法中效果最优。

(a) Algorithm of literature[15]

(b) Algorithm of literature[9]



(c) This paper

图 7　3 种算法提取的手语“松”的关键帧

Fig. 7　Key frames extracted by three algorithms of sign language “loose”

4　结束语

本文提出了一种面向手语识别的视频关键帧提取和优化算法。利用卷积自编码器提取出的图像高级特征进行 K-means 聚类, 根据清晰度算法选择聚得每一类中相似的视频帧中最清晰的一帧, 再利用手语动作中关键手势时的停顿, 通过点密度筛选出密集度大的视频帧。该算法有效地将 K-means 算法中存在的不足进行了改进, 避免了聚类算法中关键帧提取不清晰而造成的关键帧提取效果不理想的问题, 使提取的关键帧质量更高, 进而提高后期的手势识别准确率。经过实验验证, 本文算法在手语识别上具有非常快的识别速度和较高的准确率, 尤其在减少冗余帧方面表现更加出色。在今后的学习研究中, 应继续对本算法加以改进, 以提升运算速度和效率。

参考文献:

[1] CAO J, YU L, CHEN M, *et al*. A key frame selection algorithm based on sliding window and image features [C]//2016 International Conference on Parallel and Distributed Systems (ICPADS). Wuhan, China: IEEE, 2016: 956-962.

[2] CHEN L, WANG Y. Automatic key frame extraction in continuous videos from construction monitoring by using color, texture, and gradient features[J]. Automation in Construction, 2017, 81: 355-368.

[3] IOANNIDIS A, CHASANIS V, LIKAS A. Weighted multi-view key-frame extraction[J]. Pattern Recognition Letters, 2016, 72: 52-61.

[4] DEVANNE M, WANNOUS H, BERRETTI S, *et al*. 3-D Human action recognition by shape analysis of motion trajectories on riemannian manifold[J]. IEEE Transactions on Cybernetics, 2015, 45(7): 1340-1352.

[5] 马楠, 石祥滨, 代钦, 等. 一种音乐舞蹈视频关键帧提取方法[J]. 系统仿真学报, 2018, 30(7): 384-390.

[6] 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[J]. 计算机科学, 2017, 44(2): 188-193.

[7] NASREEN A, ROY K, ROY K, *et al*. Key frame extraction and foreground modelling using K-means clustering

[C]// 2015 7th International Conference on Computational Intelligence, Communication Systems and Networks(CICSyN). USA: IEEE, 2015, 34: 141-145.

[8] GHARBI H, BAHROUN S, MASSAOUDI M, *et al*. Key frames extraction using graph modularity clustering for efficient video summarization [C]// IEEE International Conference on Acoustics. USA: IEEE, 2017: 1502-1506.

[9] 赵洪, 宣士斌. 人体运动视频关键帧优化及行为识别[J]. 图学学报, 2018, 39(3): 86-92.

[10] THEIS L, SHI W, CUNNINGHAM A, *et al*. Lossy image compression with compressive autoencoders [EB/OL]. arXiv, [2017-03-01], [2019-12-01]. arXiv.1703.00395, 2017.

[11] MASCI J, MEIER U, CIRESAN D, *et al*. Stacked convolutional auto-encoders for hierarchical feature extraction [C]// 21st International Conference on Artificial Neural Networks. Espoo, Finland: Springer-Verlag, 2011: 52-59.

[12] ZENG K, YU J, WANG R, *et al*. Coupled deep autoencoder for single image super-resolution[J]. IEEE Transactions on Cybernetics, 2017, 47(1): 27-37.

[13] MENTZER F, AGUSTSSON E, TSCHANNEN M, *et al*. Conditional probability models for deep image compression[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. arXiv: 1801.04260, 2018.

[14] 杨斯涵. 基于边缘特征的单帧图像清晰度判定[J]. 计算机工程与应用, 2009, 45(30): 198-199.

[15] YANG S, LIN X. Key frame extraction using unsupervised clustering based on a statistical model[J]. Tsinghua Science and Technology, 2005, 10(2): 169-173.

Video Key Frame Extraction and Optimization Algorithm for Sign Language Recognition

ZHOU Zhou, HAN Fang, WANG Zhijie

（School of Information Science and Technology, Donghua University, Shanghai 201620, China）

Abstract: Sign language recognition technology based on computer vision brings great convenience to bilingual teaching in deaf schools, and one of the challenges of sign language recognition technology is to extract keyframes of the video. Based on the characteristics of key frames of sign language video and the signing habits of sign language speakers, this paper proposes a video key frame extraction and optimization algorithm for sign language recognition. Firstly, the convolutional auto-encoder is used to extract the deep features of video frames, which are clustered by K-means. Among each group of video frames, the clearest video frames are selected as firstly extracted keyframes by definition. Secondly, the point density method is used to optimize the extracted keyframes, and then the final key frames are obtained for gesture recognition. The experimental results show that the algorithm could reduce substantial redundant frames, and improve the accuracy and efficiency of sign language recognition.

Key words: sign language recognition; video key frame; convolutional auto-encoder; K-means; point density