基于关键帧的连续手语语句识别算法研究

郭鑫鹏[1] 黄元元[1] 胡作进[2]
(南京航空航天大学计算机科学与技术学院 南京 210016)[1]
(南京特殊教育师范学院数学与信息科学学院 南京 210038)[2]

摘 要 目前,对于动态手语的识别大多只是针对手语词汇的,对连续的手语语句的识别研究以及相应成果较少,原因在于难以对其进行有效的分割。提出了一种基于加权关键帧的手语语句识别算法。关键帧可以看作是手语词汇的基本组成单元,根据关键帧即可得到相关词汇,并将其组成连续的手语语句,从而避免了对手语语句直接做分割的难点。借助于体感设备,首先提出了一种基于手语轨迹的自适应关键帧提取算法,然后根据关键帧包含的语义对其进行加权处理,最后设计了基于加权关键帧序列的识别算法,得到连续的手语语句。实验证明,设计的算法可以实现对连续手语语句的实时识别。

关键词 手语语句识别,关键帧,手语轨迹,体感设备
中图法分类号 TP391 **文献标识码** A

Research on Continuous Sign Language Sentence Recognition Algorithm Based on Key Frame

GUO Xin-peng[1] HUANG Yuan-yuan[1] HU Zuo-jin[2]
(Institute of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)[1]
(College of Information Science, Nanjing Normal University of Special Education, Nanjing 210038, China)[2]

Abstract At present, most of the dynamic sign language recognition is only for sign language words. The continuous sign language sentence recognition research and the corresponding results are less, because the segmentation of such sentence is very difficult. In this paper, a sign language sentence recognition algorithm was proposed based on weighted key frames. Key frames can be regarded as the basic unit of sign word, therefore, according to the key frames, we can get related vocabularies, and thus we can further organize these vocabularies into meaningful sentence. Such work can avoid the hard point of dividing sign language sentence directly. With the help of Kinect, i. e. motion-control device, a kind of self-adaptive algorithm of key frame extraction based on the trajectory of sign language was brought out in the paper. After that, the key frame was given to weight according to its semantic contribution. Finally, the recognition algorithm was designed based on these weighted key frames and thus got the continuous sign language sentence. Experiments show that the algorithm designed in this paper can realize real-time recognition of continuous sign language sentences.

Keywords Sign language sentence recognition, Key frame, Gesture trace, Kinect motion-control device

1 引言

手势语言是人体语言中非常重要的一个部分,选择手语作为人机交互的语言更加直观、生动、形象。手语是聋哑人之间进行思想交流的语言,也是他们与外界进行沟通的重要方式,但目前国内专门从事手语教学的机构很少,这为聋人与健听人之间的交流造成了巨大的障碍。因此研究手语识别技术不仅可以帮助聋哑人更好地融入到社会中,而且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。

基于视觉的手语识别是一个极富挑战性的研究课题,目前已经取得不少的研究成果,但大部分是针对手语词汇的识别。例如,Starrier 等人[1]利用 HMM 算法对美国手语进行识别,识别率达到 92%;Ulrich Agris 等人[2]利用隐马尔可夫模型并结合面部特征来对手语小词汇进行识别,取得了不错的实验成果;Grobel K 和 Assam M[3]采用 HMM 技术实现了 262 个词汇的识别,正确率达到 91%;Ravikiran J 等人[4]在边界跟踪和手指检测的基础上对手指表示的字母进行识别。但是单目摄像头由于无法对手部进行精确定位,因此只能用来识别相对简单的手语词汇;而双目摄像头虽然能够获取手部的三维信息,但是在每次使用时都需要做校准,非常不方便。随着体感设备 Kinect 的推出,手语识别的研究又迈进了一个新的台阶,它不仅价格低廉而且可以检测人体 25 个关节点的数据,使用方便且无需校准,成为许多手语识别研究者的首选。它提供的骨骼数据流和距离传感器使得手部位置的检测和跟踪不再成为难题[5-7]。利用 kinect 可以识别出一些简单的手语词汇,例如文献[8]采用 K 曲率算法将 Kinect 深度信息和骨骼信息相结合来对手指尖点进行检测,平均正确率为 93%;高文等人[9]提出一种与手语者无关的手语识别研究方

本文受扬州市"绿杨金凤"优秀博士项目,江苏省"双创"项目资助。
郭鑫鹏(1992-),男,硕士生,主要研究方向为模式识别、图像处理;**黄元元**(1975-),女,博士,副教授,主要研究方向为多媒体技术、图像处理、模式识别,E-mail:805861040@qq.com;**胡作进**(1965-),男,博士,教授,主要研究方向为数据处理、机器学习。

法；Nasri 等人[10]采用基于手部图像轮廓相似性的算法对美国手语的识别率达到 90.4%；文献[11]利用手语轨迹的匹配策略对 370 个日常词汇的识别率为 78.3%；文献[12]提出一种基于 HMM 模型对手势轨迹建模的新思路，比用传统的坐标特征建模取得了更好的性能。

以上都是针对手语词汇的研究，而在连续的手语句子识别方面的研究相对较少，早期 Starner 与 Pentland 等人利用 HMM 并借助一台彩色摄像机识别出了美国手语中 40 个标有词性的词汇所组成的简单短句子[13]。方高林等人利用精简循环网(Simple Recurrent Network，SRN)/HMM 将连续手语分割成孤立词进行识别[14]，虽然可以识别一些连续的手语句子，但是其以提高算法复杂度和时间复杂度为代价。相比于手语词汇的识别，连续的手语语句的识别更难，因为目前对于自然语句的识别大多是基于 HMM 模型的。但要使用 HMM 模型对自然语句进行识别，首先需对句子做分词处理，即分割出自然语句中所包含的全部可能的词汇，然后依据 HMM 模型将这些词汇组织为有意义的句子，从而实现对自然语句的识别。但对于连续的手语语句，由于手语动作是连贯的，组成句子的词汇之间在手语动作上或者是手型变化上都无任何间隔，因此对手语语句进行有效分割一直是一个难点。无法准确分割出词汇，就很难实现对连续语句的识别。

因此，本文提出了一种基于加权关键帧的连续手语语句识别算法。一个词汇由若干关键动作组成[15]，这些关键动作被称为关键帧。如果能将这些关键动作从手语中分离出来，即可实现从连续手语中分割出单独的词汇；此外，由于每个关键动作包含的语义内容不同，因此可以根据关键动作为语义的重要性赋予一定的权值，这样，只要那些权值较大的重要的关键动作能够被准确分离并识别出来，那么对应的词汇就能被识别出来，进而可以确保连续的手语句子的语义在一些权值较低的关键动作未被准确识别的情况下依然完整，不会有缺失。

2　关键帧

一个手语动作持续若干秒后会包含若干帧图像。但其中的每一帧对手语语义的贡献是不同的。将关键手势所在的帧记作关键帧，其余不包含语义特征的帧仅仅是一些过渡帧而已，过渡帧对语义的表达无任何意义。以手语词汇“旅途”为例，手语者一的手语动作“旅途”持续 1.28 秒，按照体感设备 Kinect 30 帧/秒的采样频率，会得到 38 帧图像。但其中真正能表达“旅途”这个词的语义的只有如图 1 所示的两个关键动作，除此之外，绝大多数的帧都是过渡帧，并不包含任何语义特征。

图 1　手语者一的“旅途”关键帧

通过对大量的手语词汇的观察和统计发现，大多数的手语词汇一般仅包含 1～3 个关键动作，最多也不超过 5 个关键动作[15-16]。在一个包含几十帧甚至上百帧的手语词汇视频中，真正包含语义特征的不过几帧而已，这几帧可被称为关键帧；此外，由于手语主要靠手型的变化来表达语义，因此虽然不同的手语者在手语动作上存在较大差异，例如手部运动轨迹会因动作幅度不一样而存在差异，过渡手型有的张开或有的握拳等，但是对于相同词汇，不同手语者的关键手型是基本一致的。因此，可以将关键帧看作是手语词汇的基本组成单元。利用关键帧来描述手语词汇，一方面可以大幅度地减少数据量，另一方面也能保证对手语词汇描述的准确性与稳定性。如何提取关键帧将是手语识别的关键问题。

3　算法设计

借助于体感设备，本文提出了一种自适应的基于手语轨迹的关键帧提取算法。

3.1　基于手语轨迹的关键帧提取

Kinect 可以检测手心位置，该位置是一个包含深度信息的三维坐标。Kinect 的帧频是 30 帧/秒，将每一帧中的手心位置看作是一个采样点，将这些采样点连接起来，即可得到手语的轨迹曲线。

虽然每个人的手语习惯都不同，例如手速不同、臂展不同等，这些都将导致对于相同的手语动作，不同的人的手部的轨迹曲线会有较大差别，即关键帧的出现位置在不同的轨迹曲线上都是不一样的，但是经过大量的统计实验发现，由于关键帧包含词汇的重要语义，因此手语者在作出手语动作时，为了强调手语语义，在作出关键动作时手部往往会有一个下意识的短暂停留，并体现在手语轨迹曲线上，从而导致在空间某处位置附近采样点出现得比较密集。基于此，本文提出了一种基于手心点密度的关键帧提取算法。

首先，计算手语轨迹中各个采样点在连续时间内的点密度。定义手语轨迹曲线 ρ 上第 j 个采样点 X_j 的密度为：

$$Density(X_j)=\{X_i \mid dis(X_j,X_i)<\delta, X_i\in\rho\} \tag{1}$$

其中，X_i 表示曲线 ρ 上的第 i 个采样点，X_i 与 X_j 一样，都是一个三维的坐标向量；$dis(X_j,X_i)$ 表示 X_j 与 X_i 之间的欧氏距离；上式用于统计在曲线 ρ 上有多少个采样点与 X_j 之间的距离小于阈值 δ，满足条件的采样点 X_i 越多，表示 X_j 的点密度越大，反之则表示 X_j 的点密度越小。定义阈值 δ 为轨迹曲线上所有相邻点之间的距离之和的平均值。

$$\delta=\frac{\sum_{i=1}^{L-1} dis(X_i,X_{i+1})}{L-1} \tag{2}$$

其中，L 表示曲线 ρ 上的采样点的个数。鉴于时间的连续性，要求式(1)中满足条件的 X_i 的下标号均是连续的。

以“你去那边帮忙”手语为例，由于要考虑左右两只手的轨迹，因此首先分别计算左右两只手的轨迹点密度，再将两只手的轨迹点密度相加，从而得到手语的整体轨迹点密度，如图 2 所示。该手语动作持续 6.9 秒，一共获得 207 帧，图 2 中横坐标表示帧的编号，纵坐标表示对应帧的点密度。在得到手语轨迹的点密度曲线后，设定一个阈值 T，在理论上可以认定点密度值大于 T 的帧为关键帧。但是对于图 2 所示的手语轨迹点密度，若采用自适应的最大类间方差法求取阈值 T，如图 2 中的红线位置，那么得到的关键帧数量非常多(有 58 帧)，虽然包括了所有关键帧，但是由于动作的连贯性，很多帧

的动作是完全一样或者相似的，也包含了一些过渡帧。因此，直接对轨迹的点密度曲线做阈值分割，提取关键帧的效果并不理想。

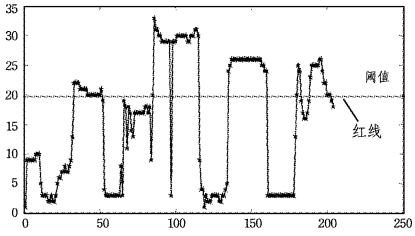


图 2 手语“你去那边帮忙”的轨迹点密度曲线

若关键帧检测不准确，则会直接影响后续的手语识别。因此采用区域划分的方法对手语轨迹的点密度曲线做处理。关键帧是手语词汇的基本组成单元，一个手语词汇包含的关键帧最多不超过 5 帧，而一个手语词汇动作一般在 3 秒内即可完成，手语语句由手语词汇构成，因此对手语轨迹的点密度曲线按照 0.5～0.8 秒的间隔进行等间隔划分，在每个区间中认定大于阈值 T 的最大值对应的点为关键帧。从而一个区间最多只有一个关键帧，区域划分的间隔保证我们可以得到实际关键帧数量 1.5～2 倍的关键帧，确保不会漏帧。如图 3 所示，对图 2 中的轨迹点密度曲线按照0.7秒的间隔进行区域划分，一共划分出 10 个间隔，每个间隔包含 20 帧。在每个区域间隔中找到点密度大于阈值的最大值对应的帧，一共找到 8 个关键帧，如图 4 所示。

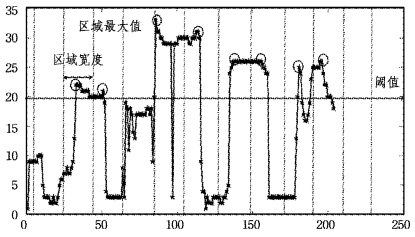


图 3 对手语轨迹的点密度曲线做区间划分处理

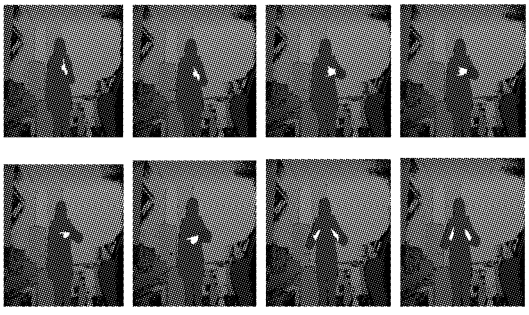


图 4 “你去那边帮忙”对应的关键帧序列

从图 4 可以看出，虽然关键帧的数量减少了，但由于动作的连贯性，还是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 4 中完整地包含了“你去那边帮忙”的所有关键帧。重复的关键帧在后续的识别中是可以被剔除的。因为正常情况下连续的关键帧是不可能重复的，所以只要在识别时发现连续的关键帧属于同一类别，即可作出剔除的处理。

该算法是在对不同的手语者做大量实验后发现的一个统计规律，该规律也符合一般人在做手语表达时的正常心理感受——明确地展示手语的语义。因此在做关键动作时，手语者都会下意识地停顿以强调语义。这种停顿并不会发生在过渡动作上，因为过渡动作并不包含语义，所以手语者的手速不会影响该算法对关键帧的提取。若手语者对手语不熟练，中途停顿下来，则视其为失败的动作，不予识别。

3.2 关键帧加权

关键帧是手语词汇的基本组成单元。但是每个关键帧对词汇的语义贡献是不同的，即每个关键帧对词汇的重要性不同。因此，可以根据关键帧对词汇语义贡献的大小来对关键帧做加权处理。那么，如何衡量一个关键帧对词汇语义的重要性呢？图 5－图 7 示出了 3 个词汇对应的关键帧，图 5、图 7 的第一个关键帧以及图 6 的第二个关键帧相同，说明这个关键帧对于这 3 个词来说辨识度不高；而图 7 中的第二个关键帧只在“旅客”这个词汇中出现，在其他词汇里中未出现，那么说明该关键帧对“旅客”这个词汇的辨识度很高。辨识度高的关键帧显然包含了词汇的重要语义，因此应该给予其较高的权值；而对于辨识度低的那些关键帧，应该给予其较小的权值。

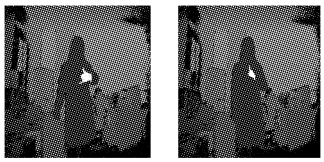


图 5 “访问”关键帧

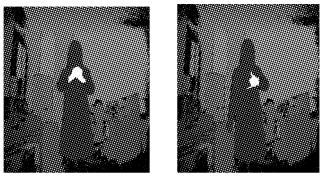


图 6 “拜访”关键帧

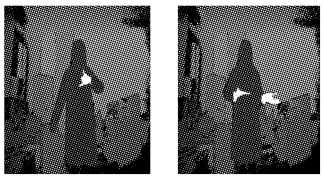


图 7 “旅客”关键帧

假定当前总共能够识别 M 个手语词汇，这些词汇一共包含 N 个关键帧，那么可以建立一个 $M\times N$ 的数组 A，其中数组元素 $a_{ij}=1$ 表示第 i 个词汇中包含第 j 个关键帧；若 $a_{ij}=0$，则表示第 i 个词汇中不包含第 j 个关键帧。将数组中每一列的元素相加，即可得到对应的关键帧在所有 M 个词汇中出现的频次 $P(j)$，该频次越大，说明该关键帧的辨识度越低，反之亦然。

$$P(j)=\frac{\sum_{i=1}^{M}a_{ij}}{\sum_{i=1}^{M}\sum_{k=1}^{N}a_{ik}},1\leqslant j\leqslant N \tag{3}$$

有的关键帧只在一个词汇中出现，有的关键帧会在多个词汇中出现，相同的关键帧在不同词汇中的作用是不同的，因此为每个关键帧在所有词汇中分别赋予权值。再创建一个 $M\times N$ 的数组 W，其元素 w_{ij} 表示第 j 个关键帧在第 i 个词汇中的权值，其表达式如下：

$$w_{ij}=\frac{a_{ij}/P(j)}{\sum_{k=1}^{N}(a_{ik}/P(k))} \tag{4}$$

这样可以保证出现频次高的关键帧在对应词汇中的权值较小，而出现频次较低的关键帧在对应词汇中的权值则较大，同时保证一个词汇中包含的所有关键帧，其权值累加和为 1，即：

$$\sum_{j=1}^{N}w_{ij}=1 \tag{5}$$

3.3　加权关键帧序列的识别

对提取到的关键帧中的关键动作（即关键手型）作分类识别。首先利用 Kinect 检测的手心位置，以及手心位置处的深度信息分割出手部区域，然后对手部区域提取特征、设计分类器等。这里对关键帧的分类识别不做详述，因为手型图像属于二值化的简单图像，对这样的手型进行分类识别是完全可以实现的[17]。

一个手语句子由若干个手语单词组成，一个手语单词又由若干个关键帧组成，因此一个手语句子是由一组关键帧组成的，对句子的识别就转化成对这一组关键帧进行识别。对于不同的单词，关键帧对应不同的权值，因此对关键帧序列进行识别的基本思想就是找出累加权值最大的关键帧组合，这些组合即为构成句子的单词，连接这些单词即可完成连续的手语语句的识别。

假定对一个手语语句提取关键帧并做识别后，去除重复的关键帧和非关键帧，得到最终的关键帧序列 $K_1,K_2,\cdots,K_L$，其中 L 是关键帧的个数。目前系统中包含的关键帧总数为 N，可识别的词汇数是 M，由于每个手语词汇中包含的关键帧数量最多不超过 5 个，因此对上述关键帧序列的识别过程为：

(1)以起始的连续 5 帧 K_1-K_5 为一组，每一帧在矩阵 W 中对应一列，在 W 的每一行中，将 K_1-K_5 对应的列元素累加。W 中的每一行对应一个词汇，因此哪一行的累加和最大，说明这一组关键帧对应的就是哪个词汇。但是，有可能会出现多行的累加和相等并都是最大值，此时应该选取包含关键帧数量最多的词汇。

强调包含关键帧数量最多的词汇的原因在于有的词汇的关键帧具有一定的包含关系，如图 8 所示的两个关键帧。两个关键帧对应“旅途”词汇，在“旅途”中关键帧 1 的权值为 0.33，关键帧 2 的权值为 0.67；但是关键帧 1 本身又对应“去”这个单词，由于“去”中只有这一个关键帧，因此在“去”中关键帧 1 的权值最大，为 1。如果以关键帧 1 为起始帧来识别某个词汇时“去”和“旅途”的累加权值都为 1，那么应该识别哪个词汇呢？该情况一般需选择包含关键帧数量最多的单词，即识别“旅途”，而不是“去”。同样的情况还有“我”和“我们”，“你”和“你们”等词汇。

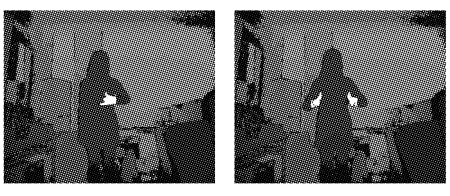


图 8　关键帧序列示例

(2)通过上述步骤识别第一个词汇 t_1 后，很有可能不是所有 K_1-K_5 序列中的关键帧都对该单词集合 t_1 有语义贡献，即有的关键帧对于 t_1 的权值为 0。由此，找到在时间上最早出现且对当前 t_1 的权值为 0 的关键帧 K_e，其中 $2\leqslant e\leqslant 5$，且在 K_e-K_5 之间的关键帧对当前 t_1 的权值均为 0。以 e 作为下一组关键帧的起始帧，再对 K_e-K_{e+4} 这 5 个关键帧序列重复上述步骤，从而得到下一个单词 t_2；一直重复该过程，如果当前的起始帧到最后一帧 K_L 之间不足 5 帧，继续按照上述的步骤进行处理，直到最后一帧 K_L 对当前获取的单词 t_q 的权值不为零为止。

(3)有可能出现如下情况：在某次识别过程中以 K_e 作为起始帧，在 K_e-K_{e+4} 这 5 个关键帧中识别某个词汇时，以 K_e 作为起始帧的最大累加权值未超过 0.5，则说明当前的 K_e 极有可能是一个错帧，将其从关键帧序列中删除，从 K_{e+1} 开始继续重复执行上述步骤。

根据上述步骤，可以对关键帧序列进行识别，得到按时间顺序最有可能出现的若干词汇 $t_1,t_2,\cdots,t_q$，连接这些词汇即可得到相应的连续语句。由于对关键帧做了加权处理，因此当关键帧的提取或者识别出现漏帧或者错帧时，也有可能得到正确的识别结果。

4　实验数据与分析

Kinect 对硬件环境的要求较高，本文的硬件环境为：CPU：Inter（R） Core（TM） i7-4790 CPU @ 3.60GHz 3.60GHz；安装内存（RAM）：8.00GB。操作系统为 Windows 10 专业版 64 位，在 Kinect Studio V2.0 中利用 C# 作为开发语言进行实验。

为了验证本文算法的有效性，进行了两类实验，是验证关键帧提取算法的准确性与稳定性，其次验证加权的关键帧对手语语句识别的作用。本文选择了 26 个常用词汇，对这 26 个词汇定义关键帧并对关键帧做加权处理。这 26 个词汇一共包含 46 个关键帧，理论上均可对这 26 个词汇所组成的语句作出识别。

4.1　关键帧提取实验

本文选择了 5 位手语者，其中前 3 人为较熟练的手语者，后 2 人现场学习手语，现场拍摄。利用系统的 26 个词汇组成 10 个不同的句子，让 5 位手语者对这 10 个句子分别表演 3 遍，从而共得到 150 个手语样本。对这些样本分别计算手语轨迹的点密度，再根据点密度曲线提取关键帧。可以发现，现场学习手语的手语者们的点密度曲线在关键帧位置处尤其明显，即关键帧处的轨迹点密度特别大。因为初学者动作不熟练，对重点动作（即关键动作）尤其想重点突出，所以会在此处刻意停留，从而造成此处的轨迹点密度明显较大；而由于其他 3 位熟练的手语者的动作连贯性更好，因此在关键帧处不会刻意停留，只是动作表达的一种下意识停顿，故他们的手语轨迹在关键帧处的点密度不如初学者的大，但是也比过渡动作处的点密度大。表 1 列出了对这 5 位手语者提取关键帧的漏帧率。关键帧提取的关键在于不能漏帧，漏帧意味着语义的缺失；而若有重复关键帧或者非关键帧被提取，则可以通过分类器或者后期的识别来进行调整。由表 1 可以看出，本文所提出的关键帧提取算法的漏检率较低，算法简单且容易实现，效率较高。

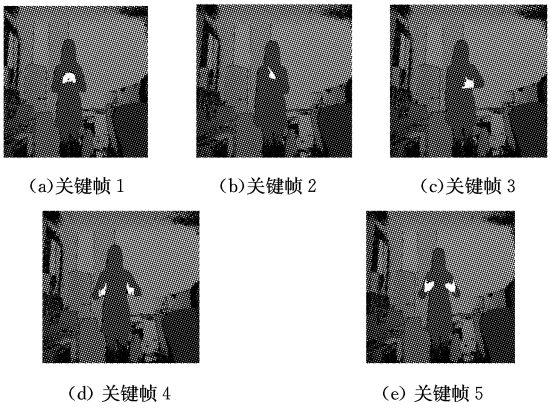
表 1 关键帧提取的漏检率

手语者编号	1	2	3	4	5
漏帧率值	0.010	0.0108	0.0195	0	0

4.2 手语语句识别实验

对上述手语者作出的手语动作提取关键帧后，假定事先通过样本采集、特征提取以及大量的训练设计出分类器，对提取到的关键帧序列逐一做分类识别，剔除重复的关键帧以及非关键帧，从而得到最终的关键帧序列。图 9 给出了针对某个手语者的手语动作“祝你旅途愉快”得到的最终关键帧序列。

由于系统中包含 26 个词汇，这 26 个词汇由 46 个不同的关键帧组成，因此对于图 9 所示的关键帧序列，识别结果分别是关键帧的第“19”“10”“12”“7”和“15”类别。这些关键帧在不同词汇中的权值如表 2 所列。



(a)关键帧 1 (b)关键帧 2 (c)关键帧 3

(d) 关键帧 4 (e) 关键帧 5

图 9 “祝你旅途愉快”关键帧序列

表 2 关键帧在词汇中的权值分配(部分)

词汇 \ 关键帧类别编号	··	7	··	10	11	12	··	15	··	19	··
你	…	0.0	…	1.0	0.0	1.0	…	0.0	…	0.0	…
去	…	0.0	…	0.0	0.0	0.0	…	0.0	…	0.0	…
祝	…	0.0	…	0.0	0.0	0.0	…	0.0	…	1.0	…
旅途	…	0.67	…	0.0	0.0	0.33	…	0.0	…	0.0	…
愉快	…	0.0	…	0.0	0.0	0.0	…	1.0	…	0.0	…
健康	…	0.0	…	0.0	0.36	0.0	…	0.0	…	0.0	…
。。	…	…	…	…	…	…	…	…	…	…	…

由于系统中包含的 26 个词汇的关键帧数量都不超过 3，因此在利用 3.3 节中设计的算法识别关键帧序列时，选择步长为 3，即每次以 3 个关键帧为一组来识别对应的词汇。以图 9 所示的关键帧序列为例，此时的关键帧序列为“19”“10”“12”“7”和“15”，识别过程如下：

(1)首先以关键帧“19”为起始帧，在“19”“10”“12”这 3 个关键帧中确定第一个词 t_1。在表 2 中，包含关键帧“19”、累加和最大且关键帧数量最多的词为“祝”，因此 t_1＝“祝”。

(2)然后以“10”为起始帧，在“10”“12”“7”中确定第二个词 t_2。在表 2 中，包含关键帧“10”、累加和最大且关键帧数量最多的词是“你”，因此 t_2＝“你”。

(3)再以“12”为起始帧，在“12”“7”和“15”中确定第三个词 t_3。在表 2 中，包含关键帧“12”、累加和最大且关键帧数量最多的词是“旅途”，因此 t_3＝“旅途”。

(4)目前只剩下关键帧“15”，在表 2 中其对应的词汇为“愉快”，因此最后一个词 t_4＝“愉快”。

(5)关键帧识别结束，将得到的词汇连起来，从而得到连续的句子“祝你旅途愉快”。识别结束。

由上述过程可知，在关键帧序列识别准确的情况下，根据所设计的算法，可以由关键帧序列得到连续的、语义完整的手语句子。

但如果关键帧不完整或者识别有误，那么由于关键帧带有权值，只要那些重要的、辨识度高的关键帧被正确识别出，就有可能得到正确的识别结果。例如，在上述的关键帧序列中图 9(c)这个关键帧在提取时被漏掉了，那么实际得到的关键帧序列则为“19”“10”“7”和“15”，对这一序列进行识别时，t_1 和 t_2 不变，但是在识别 t_3 时，包含关键帧“7”的累加和最大只有 0.67，虽然未达到 1，但是也超过了我们设定的阈值0.5，因此“旅途”这个词还是可以被识别。因此在关键帧缺失的情况下，依然可以得到正确的识别结果。

再比如，图 9(c)这个关键帧在识别时发生错误分类，将其归为关键帧第“11”个类别，那么实际得到的关键帧序列为“19”“10”“11”“7”和“15”。对这一序列进行识别时不会对 t_1 和 t_2 的识别产生影响。在“11”“7”和“15”中识别 t_3 时，包含“11”的最大累加和只有 0.36，该值未超过阈值 0.5，因此可以判定当前的起始帧“11”是一个错帧，并将其从序列中删除，继续以下一个关键帧“7”为起始帧识别词汇 t_3。由于关键帧“7”包含了词汇“旅途”的重要语义，对于旅途有较高的辨识度，因此“旅途”这个词能够被识别，最终可以得到正确的识别结果。

但是如果包含重要语义的关键帧被漏检或误识，那么很可能导致语义的缺失，从而使识别结果不够准确。因此，本文设计的算法对关键帧的提取以及识别要求较高。

本文设计的实验系统可以对 26 个常用词汇的关键帧做出识别，那么理论上本系统均可实现识别由这 26 个词汇组成的句子。我们依然选择关键帧提取实验中的 5 位手语者以及他们所做的 150 个连续手语句子作为实验对象，识别情况如表 3 所列。由表 1 可知，对这 5 个人提取关键帧时准确率较高，漏帧率较低，表明关键帧基本全部被分离，保证了语义的完整性；最终的识别准确率主要依靠对关键帧的识别。由于本文的研究重点在于关键帧的提取以及如何利用关键帧识别连续语句，不在于如何对关键帧做分类和识别。因此实验中对关键帧的识别采用的是文献[17]中的随机森林，其中前两个手语者的关键动作作为训练样本设计随机森林，后 3 个手语者的动作未经过训练。表 3 中，后两位手语者的识别准确率不高，主要原因在于他们均是现场学习手语，动作不够规范；虽然第三位手语者的动作未加入训练集，但是对他的识别准确率较高，因为他是一位熟练的手语者，动作比较规范；前两位的识别率最高的原因在于：1)他们的动作被加入了训练集；2)他们都是动作较为规范的手语者。

表 3　不同手语者的手语识别结果

手语者编号	1	2	3	4	5
识别准确率	0.9933	0.9777	0.9400	0.8067	0.7500
识别所用时间/ms	985	1073	1129	1079	1089

从整体来说，本文设计的算法的可行性较高，1 秒内即可得到识别结果，实时性可以得到保障。此外，Kinect 本身对使用环境没有特殊要求，仅要求手语者与镜头之间保持0.8～1米的距离，且之间无任何障碍物即可。因此本系统在使用时不受外界环境、光照等影响。

结束语　本文提出了一种对连续的手语语句进行有效识别的新方法，即利用手语者的心理暗示，根据手语轨迹提取关键帧，利用关键帧描述手语词汇，再利用词汇描述连续句子，同时利用关键帧的权值提高识别的准确率。该算法对于不同的手语者（无论是初学还是相对较熟练的人）均取得了较好的效果。该算法的实现依赖于关键帧的提取与识别，因此为了更为稳定地提取关键帧，除了需考虑点密度的变化，还应该加入对手型变化的考虑。此外，在提取关键帧的基础上获得准确、稳定的手语识别结果，依赖于对关键帧的正确识别与分类，因此对关键手型的特征提取以及分类器设计提出了更高的要求。这些都是今后需要进一步深入研究的问题。

参 考 文 献

[1] STARNER T, PENTLAND A. Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video[C]// IEEE Transaction on Pattern Analysis and Machine Intelligence. 1998: 1371-1375.

[2] VON AGRIS U, ZIEREN J, CANZLER U, et al. Recent developments in visual sign language recognition[J]. Universal Access in the Information Society, 2008, 6(4): 323-362.

[3] GROBEL K, ASSAN M. Isolated sign language recognition using hidden Markov models[C]// IEEE International Conference on Systems. 1997: 162-167.

[4] RAVIKIRAN J, MAHESH K, MAHISHI S, et al. Finger Detection for Sign Language Recognition[J]. International Association of Engineers, 2009, 2174(1): 489-493.

[5] GUO X L, YANG T T. Gesture recognition based on HMM-FNN model using a Kinect[J]. Journal on Multimodal User Interfaces, 2016: 1-7.

[6] 谈家谱，徐文胜. 基于 Kinect 的指尖检测与手势识别方法[J]. 计算机应用，2015，35(6)：1795-1800.

[7] HALIM Z, ABBAS G. A Kinect-Based Sign Language Hand Gesture Recognition System for Hearing- and Speech-Impaired: A Pilot Study of Pakistani Sign Language[J]. Assistive Technology, 2015, 27(1): 34-43.

[8] YAN H, ZHANG M, TONG J, et al. Real time robust multi-fingertips tracking in 3D space using Kinect[J]. Journal of Computer-Aided Design and Computer Graphics, 2013, 25(12): 1801-1809.

[9] JIANG F, GAO W, WANG C L, et al. Development in Signer-Independent Sign Language Recognition and the Ideas of Solving Some Key Problems[J]. Journal of Software, 2007, 18(3): 477-489.

[10] NASRI S, BEHRAD A, RAZZAZI F. A novel approach for dynamic hand gesture recognition using contour-based similarity images[J]. International Journal of Computer Mathematics, 2015, 92(4): 662-685.

[11] LIN Y, CHAI X, ZHOU Y, et al. Curve Matching from the View of Manifold for Sign Language Recognition[C]// Workshop on Feature and Similarity Learning for Computer Vision (FSL-CV). ACCV, 2014: 233-246.

[12] PU J, ZHOU W, ZHANG J, et al. Sign Language Recognition Based on Trajectory Modeling with HMMs[C]// International Conference on Multimedia Modeling. 2016: 686-697.

[13] STARNER T, PENTL A. Visual Recognition of American Sign Language Using Hidden Markov Models[C]// International Workshop on Automatic Face & Gesture Recognition. 1995: 189-194.

[14] 方高林，高文，陈熙霖，等. 基于 SRN/HMM 的非特定人连续手语识别系统[J]. 软件学报，2002，13(11)：2169-2175.

[15] 中国残疾人联合会教育就业部，中国聋人协会. 中国手语[M]. 华夏出版社，2003.

[16] LI S R, HUANG Y Y, HU Z J, et al. Key Frame Detection Algorithm based on Dynamic Sign Language Video for the Non Specific Population[J]. International Journal of Signal Processing: Image Processing and Pattern Recognition, 2015, 8(12): 135-148.

[17] SHI M M, HUANG Y Y, HU Z J. Dynamic Sign Language Recognition Algorithm Using Weighted Gesture Units[J]. Journal of Information and Computational Science, 2015, 12(15): 5611-5621.

（上接第 177 页）

[5] 刘威，赵文杰，李成，等. 基于改进 ORB 特征匹配的运动小目标检测[J]. 光电工程，2015，42(10)：13-20.

[6] LI J, XU T, ZHANG K. Real-Time Feature-Based Video Stabilization on FPGA[J]. IEEE Transactions on Circuits & Systems for Video Technology, 2017, 27(4): 907-919.

[7] LIANG W, XIE X, WANG J, et al. A SIFT-based mean shift algorithm for moving vehicle tracking[C]// Intelligent Vehicles Symposium Proceedings. IEEE, 2014: 762-767.

[8] 程德强，郭政，刘洁，等. 一种基于改进光流法的电子稳像算法[J]. 煤炭学院，2015，40(3)：707-712.

[9] KOH Y J, LEE C, KIM C. Video Stabilization Based on Feature Trajectory Augmentation and Selection and Robust Mesh Grid Warping[J]. IEEE Transactions on Image Processing, 2015, 24(12): 5260-5273.

[10] XUE X, CHEN Z, WANG Q, et al. Moving target detection of OFDM-MIMO radar based on ST-DFT[J]. Electronics Letters, 2015, 52(2): 150-152.

[11] K M, CHANNAPPAYYA S S. An Optical Flow-Based Full Reference Video Quality Assessment Algorithm[J]. IEEE Transactions on Image Processing, 2016, 25(6): 2480-2492.

[12] GU B, HU H, REN Y, et al. Moving Target Detection and Tracking in Complex Background[J]. International Journal of Smart Home, 2015, 9(9): 95-102.

[13] SILVA G, RODRIGUEZ P. Jitter invariant incremental principal component pursuit for video background modeling on the TK1[C]// Asilomar Conference on Signals, Systems and Computers. IEEE, 2015.

[14] 闫利，陈林. 一种改进的 SURF 及其在遥感影像匹配中的应用[J]. 武汉大学学报(信息科学版)，2013，38(7)：770-773.

[15] 程德强，郭政，刘洁，等. 一种基于改进光流法的电子稳像算法[J]. 煤炭学报，2015，40(3)：707-712.