软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(Suppl.(2)):36−43

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

在频域上利用三维轨迹匹配进行手语识别*

林宇舜[1,2], 柴秀娟[1], 许志浩[1], 尹 芳[1,2], 陈熙霖[1,2]

[1](中国科学院 计算技术研究所 智能信息处理重点实验室,北京 100190)

[2](中国科学院大学,北京 100049)

通讯作者: 陈熙霖, E-mail: xlchen@ict.ac.cn

摘 要: 手语是聋哑人互相之间常用的交流手段.但由于大部分口语使用者不懂手语, 因此影响了聋哑人参加正常的社交活动.因此,提出了一种利用简单的三维轨迹信息进行小规模手语词汇识别的方法,试图帮助聋哑人克服部分交流障碍.首先,对 Kinect 获取的三维轨迹进行预处理——对获得的三维轨迹根据打手语人的身高进行归一化,然后使用插值算法对轨迹进行均匀的指定点数的重采样.在进行匹配之前,测试集和原型图像集中的轨迹将会对齐,并使用 DFT 变换到频域空间,得到由实部、虚部、幅值串接而成的新的特征向量.最后,在频域中计算两条轨迹之间的欧氏距离以评估两条三维轨迹的相似度.对 239 个手语词汇集合的实验结果表明,该方法对于中国手语的孤立词识别是有效的.

关键词: 手语识别;三维轨迹;频域;曲线匹配

中文引用格式: 林宇舜,柴秀娟,许志浩,尹芳,陈熙霖.在频域上利用三维轨迹匹配进行手语识别.软件学报,2014,25(Suppl.(2)):36−43. http://www.jos.org.cn/1000-9825/14021.htm

英文引用格式: Lin YS, Chai XJ, Xu ZH, Yin F, Chen XL. Sign language recognition by 3D trajectory matching in frequency domain. Ruan Jian Xue Bao/Journal of Software, 2014,25(Suppl.(2)):36−43 (in Chinese). http://www.jos.org.cn/1000-9825/14021.htm

Sign Language Recognition by 3D Trajectory Matching in Frequency Domain

LIN Yu-Shun[1,2], CHAI Xiu-Juan[1], XU Zhi-Hao[1], YIN Fang[1,2], CHEN Xi-Lin[1,2]

[1](Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, The Chinese Academy of Sciences, Beijing 100190, China)

[2](University of Chinese Academy of Sciences, Beijing 100049, China)

Corresponding author: CHEN Xi-Lin, E-mail: xlchen@ict.ac.cn

Abstract: For hearing-impaired people, sign language is a common communication means just like spoken language to ordinary people. Because most of ordinary people cannot understand sign language, it's difficult for hearing-impaired community to participate in social activities. This paper proposes an effective method for sign language recognition with simple 3D trajectory information to break down the barriers between hearing-impaired and normal persons. First of all, the 3D trajectory captured from Kinect is preprocessed, and both the trajectories from the probe and galleries is normalized by the size of the signer. Then the trajectories are resampled evenly by a fast, easy, and usable interpolation algorithm. Before matching of two curves, the trajectory of the probe is aligned to the gallery trajectory and both aligned trajectories are transferred into frequency domain by DFT, and vectors linking the real part, imaginary part and amplitude are obtained. Finally, Euclidean distance between two trajectories in frequency domain is calculated to evaluate the dis-similarity of two trajectory vectors according to the minimized distance. The experimental results on a data set of 239 sign words show that the presented approach is effective to recognizing isolated words of Chinese sign language.

Key words: sign language recognition; 3D trajectory; frequency domain; curve matching

* 基金项目: 国家自然科学基金(61001193); 微软亚洲研究院项目

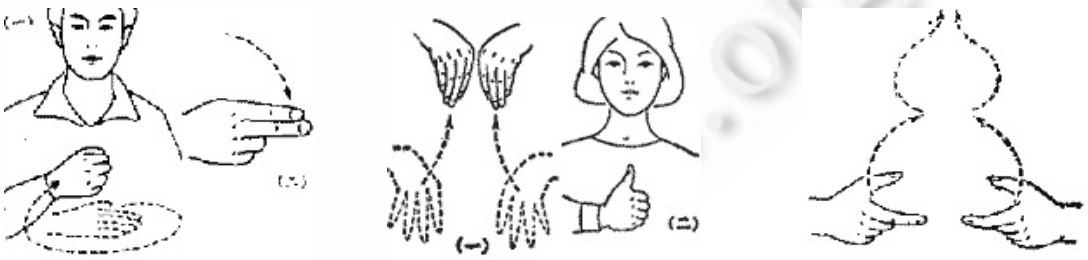
收稿时间: 2013-06-15; 定稿时间: 2013-08-21

利用手语,聋哑人互相之间能够方便地交流.但是,大部分的正常人并不懂手语.为了让聋哑人和正常人能够更好地互相交流和理解,手语识别是一项十分有益的技术.不像语音识别已经可以投入商业应用,手语识别(SLR)技术距离真正实用化还有很多的挑战,如手势不变特征的提取,手势之间过渡的模型等等.

早期的手语识别方法使用了人工神经网络(ANN).Murakami 和 Taguchi 在 1991 年使用数据手套获得的特征训练了一个 ANN[1].1995 年,Huang 等人提出了一种基于 Hopfield ANN 的孤立词手语识别系统[2].在那之后,Kim 等人在 1996 年使用数据手套提供的 x,y,z 坐标和角度,训练了一个 Fuzzy Min Max ANN,在 25 个孤立词上获得了 85%的识别率[3].由于手语识别和语音识别问题的相似性,基于 HMM 的方法在 90 年代中期后开始流行.Grobel 和 Assan 在 1997 年建立了一种基于 HMM 的孤立手语词(手势)的识别系统[4],该系统在给定的条件下表现良好.1998 年,Starner 等人的工作显示了 HMM 是手语识别的一种强有力的方法[5].

近年来,在许多领域都有与手语识别有关的工作.2007 年,Mitra 和 Acharya 介绍了用人体不同部位进行手势识别的许多方法[6].Yang 和 Sarkar 在 2008 年提出了一种称为耦合分组与匹配的方法,该方法无需对场景进行完美的分割[7],即可实现对手语词的匹配.Shi[8]等人在 2011 年提出一种直观有效的基于裸手双手跟踪的实时交互方法.除此以外,还有很多研究人员试图通过增加深度信息来帮助进行手势分析方面的工作.Zhu[9]等人融合了 Time-of-flight 深度传感器和立体摄像机两种方法获得了更高精度的深度图像.Ren[10]和 Zafrulla[11]等人尝试了将 Kinect 用于手势和美国手语的识别.不同于其他复杂而贵重的设备(如数据手套和立体摄像机),Kinect 是一种十分便宜方便的 RGB 和深度摄像机.Tong[12]等人在 2011 年提出了用 Kinect 进行三维人体扫描的方法,使用多个 Kinect 提高了精度并解决了 Kinect 之间的互相干扰的问题.2012 年,Wang 等人提出了一种名为傅里叶时序金字塔的全新时序模式表示.这种频域上的特征对于深度上的噪声是足够鲁棒的[13].

一般认为,手语词由两类最主要的特征组成——手移动的轨迹和手型.张毅[14]和邓瑞[15]等人分别利用 Kinect 获取的深度图像信息得到的轨迹和手型的特征对简单的几个手势进行了有效的识别.大部分的手语词都有着可以区分的轨迹.图 1 给出了几个标准中国手语词汇对应的图示,可见仅根据轨迹信息即可对这些手语词汇进行有效的区分.而且,在 Kinect 的帮助下,可以相对简单而精确地获得手部移动的三维轨迹.因此,本文试图基于 Kinect,利用 3D 轨迹特征进行部分手语词的识别.来自测试集的轨迹将与来自原型图像集中的轨迹对齐,然后在频域上计算不同轨迹特征向量之间的欧氏距离.



(a) 总裁 (b) 总理 (c) 葫芦

Fig.1 Pictures of three different Chinese sign language words

图 1 3 个不同的中国手语词汇示意图

本文第 1 节介绍手语识别算法的框架.轨迹匹配的算法将在第 2 节讲述.第 3 节是实验和相关结果.最后是本文的结论.

1 手语识别算法框架

本文提出的算法主要流程分为 4 个步骤.第 1 步是对输入的数据进行预处理.接着将输入轨迹和原型图像集中的轨迹进行对齐.第 3 步是使用离散余弦变换(DFT)将数据由空间域变换到频域.最后,在频域中计算出输入轨迹与原型图像集中的轨迹之间的距离.图 2 给出了算法的流程.

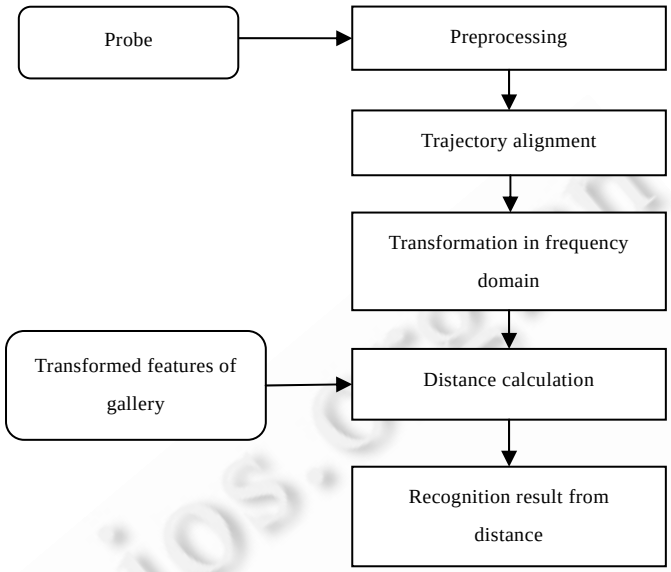


Fig.2 Framework of sign language recognition

图 2 手语识别的框架

2 基于 DFT 的 3D 手语识别

2.1 预处理

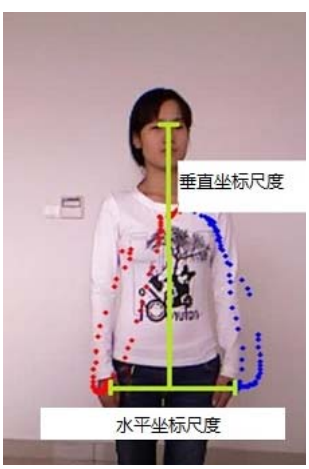
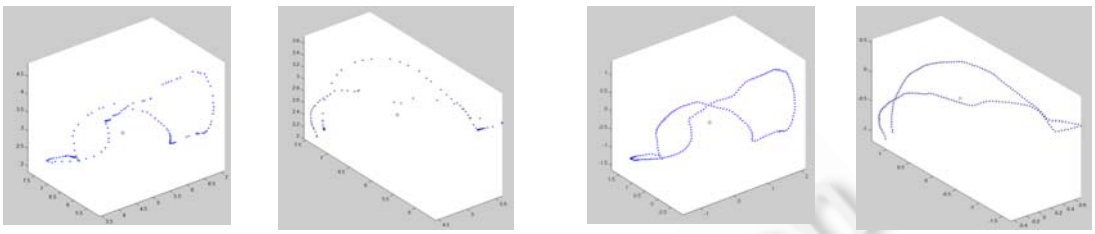


Fig.3 Normalization of trajectories

图 3 轨迹归一化图解

从 Kinect 获取的原始轨迹数据是不规则的.这是因为不同的打手语的人和 Kinect 之间的距离仅仅是大致相同的,并不精确.另一方面,打手语的人会有不同的身高和身体比例(例如,某些人的胳膊会比一般人要长一些).这会导致手语轨迹的尺度差异,从而直接匹配会产生更大的误差.为了让轨迹在一致的尺度上进行度量,首先需要根据个人的身体参数对原始轨迹数据进行归一化处理.每一个维度的单位长度将会根据打手语者的水平和垂直的坐标尺度进行归一.图 3 给出了归一化的基本原则.其中,水平的坐标尺度由第 1 帧时左手和右手的水平距离定义,垂直的坐标尺度被定义为第 1 帧时头部到左右手的垂直距离.

另外,手部的跟踪是由 Kinect 给出的 3D 骨架信息中提取出来的,并不总是稳定准确的,噪声十分常见.为了减小噪声的影响,我们使用了中值滤波.同时,为了对轨迹曲线的相似度进行评估,我们对手语词汇的整体轨迹进行了重采样,使之成为一个具有标准采样点的向量.这里我们采用了一个均匀插值的重采样算法来获得这个新的向量.首先,计算出所有 M 点的轨迹路径长度,然后将这个长度除以(N–1)来获得新的 N 个点的每一步插值的步长 I.然后轨迹会逐步延伸,当距离超过步长 I 的时候,就通过线性插值加入一个新的点[16].图 4 给出了一些轨迹重采样的示例.



(a) 原始轨迹 (b) 重采样之后的轨迹

Fig.4 Examples of resampled trajectories

图 4 重采样的轨迹示例

2.2 轨迹对齐

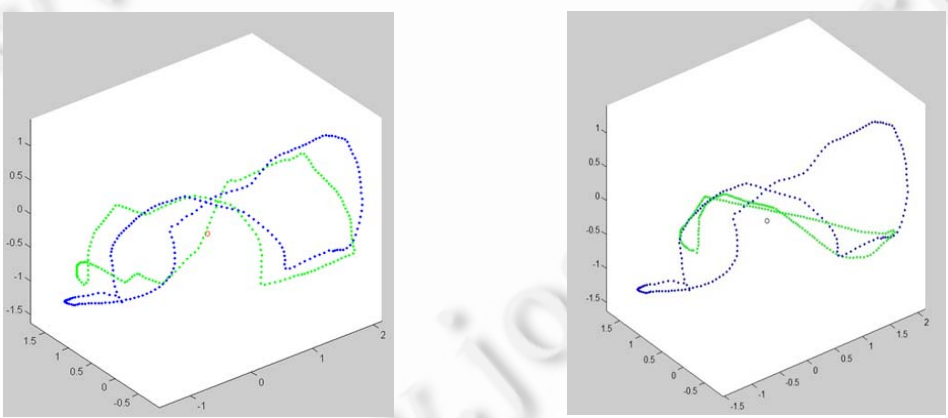
在对轨迹进行了预处理之后,还需要在匹配之前将不同的轨迹进行对齐.对齐通常包括 3 个方面:平移,旋转和放缩.由于打手语者是面对 Kinect 的,因此其角度是相对固定的,轨迹的朝向可以被认为是一个对词汇识别有用的线索.至于尺度的放缩,已经在预处理步骤的归一化过程中被考虑到了.因此,在这里的轨迹对齐部分,旋转和放缩都是不必要的,甚至是不合适的.然而,对平移的需求是显然的.尽管在预处理中已经进行了归一化处理,但是轨迹的位置依然随机地取决于打手语者.我们首先计算得到轨迹的中心 $\bar{p}(x,y,z)$:

$$\bar{p}(x,y,z)=\frac{1}{n}\sum_{i=1}^{n}p_i(x,y,z) \tag{1}$$

其中 n 是轨迹点的个数.平移之后新的点的坐标为 p_i':

$$p_i'(x,y,z)=p_i(x,y,z)-\bar{p}(x,y,z) \tag{2}$$

最终,轨迹点的中心将会被平移到原点(0, 0, 0),如图 5 所示.



(a) 相同手语词轨迹对齐 (b) 不同手语词轨迹对齐

Fig.5 Examples of trajectory alignment for the same and different sign words

图 5 相同和不同手语词汇对应的轨迹对齐示例

2.3 频域中的3D轨迹匹配

为有效的对 3D 轨迹进行匹配计算,减少轨迹噪声的影响,本文试图在频域中对 3D 轨迹信号进行分析.为将对齐后的轨迹从空间域变换到频域,本文进行了离散傅里叶变换(DFT).不同轨迹在频域上的特征可以将它们互相之间区分开来.同时,利用高频部分没有包含多少信息的特点,可以通过去除高频部分来达到减少运算和去除噪声的目的.实验结果表明在频域上的分析要比在空间域上的分析显得更稳定.

离散傅里叶变换是傅里叶变换的离散形式.它将函数上有限的样本点,转换成一系列有限的正弦曲线的复

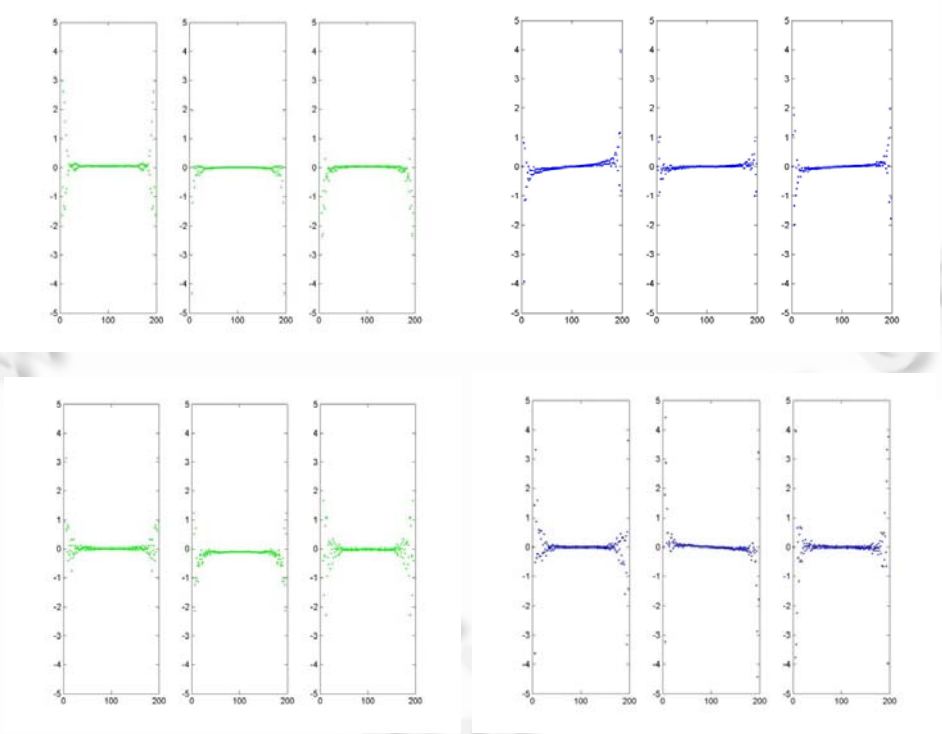
系数,这些系数按照正弦曲线的频率排列.换句话说,它将一个采样的函数从原来的域变换到了频率域.

DFT 的形式化如下所示:

$$X_K = \sum_{n=0}^{N-1} x_n \cdot \mathrm{e}^{-i2\pi kn/N} \tag{3}$$

其中, $x_n(0 \leqslant n \leqslant N-1)$ 是函数的 N 个样本点,X_K 是正弦曲线的系数.

如果将三维轨迹看作一种信号,则其每一维度可以视为信号的一个通道.该信号在每一个通道上的变化表征了不同手语词的特征.对每一个通道(维度)的向量 x,y,z 进行 DFT 变换,就可以分别获得对应的通道(维度)在频域上的系数向量 X,Y,Z(其中, $x=(x_0,x_1,...,x_{N-1})$, $X=(X_0,X_1,...,X_{N-1})$,其余同理.)在变换之后,X,Y,Z 的实部、虚部和复数模将会被抽取出来连接成轨迹的一个全新的特征向量.图 6 给出了两个不同手语词在 DFT 变换后的实部和虚部的系数,从左到右的三列分别对应 x,y,z 这 3 个通道.由图中可见,两端的低频区域包含了手语词的大部分能量,且不同的词语在实部和虚部的低频部分的系数均有较为显著的差别.因此,可以通过频域系数的差距将不同的手语词区分开来.



(a) 不同词语在 DFT 变换后实部的系数 (b) 不同词语在 DFT 变换后虚部的系数

图 6 不同手语词在 DFT 后得到的频域系数的实部(图(a))和虚部(图(b))的示例

经过 DFT 变换之后,就得到了同时包含频域的实部、虚部和复数模在 3 个通道(维度)上的系数的特征向量,该向量即可作为 3D 轨迹曲线的特征描述.因此轨迹的匹配即可转化为两个特征向量之间的匹配,并可通过不同类型的距离度量来评估,如直方图交、余弦距离等.在实现中,我们采用了最直观且简单有效的欧氏距离来评价不同特征向量之间的相似性,识别结果为具有最小距离的向量对应的词汇.

3 实验和分析

为了评估算法的性能,我们在采集的手语数据集上进行了实验.该数据集包含了 239 个中国手语词汇,每个词由聋哑学生打 5 次.数据集中的每个手语词都是从日常交流中常用的词汇里随机选取的.在实验中,5 组数据

轮流作为测试集,其他 4 组作为原型集合以进行交叉验证.每个词汇的每个样本都会和当前的测试样本计算一个距离,其中最小的距离将作为该测试样本和该词汇的最终距离.最后,与测试样本的距离最小的词汇将作为识别结果.表 1 和表 2 给出了识别率的结果,其中包括了对二维和三维的轨迹的实验.

Table 1 Recognition result of sign language based on 2D trajectory matching

表 1 基于二维轨迹匹配的手语识别结果

	P50	P51	P52	P53	P54	Average
Top 1	0.598	0.720	0.686	0.736	0.753	0.699
Top 3	0.774	0.858	0.883	0.891	0.866	0.854
Top 5	0.858	0.883	0.921	0.941	0.900	0.900
Top 10	0.908	0.954	0.954	0.987	0.946	0.950

Table 2 Recognition result of sign language based on 3D trajectory matching

表 2 基于三维轨迹匹配的手语识别结果

	P50	P51	P52	P53	P54	Average
Top 1	0.799	0.841	0.858	0.849	0.849	0.839
Top 3	0.916	0.950	0.946	0.962	0.950	0.945
Top 5	0.958	0.971	0.975	0.987	0.971	0.972
Top 10	0.975	0.992	0.992	0.992	0.987	0.987

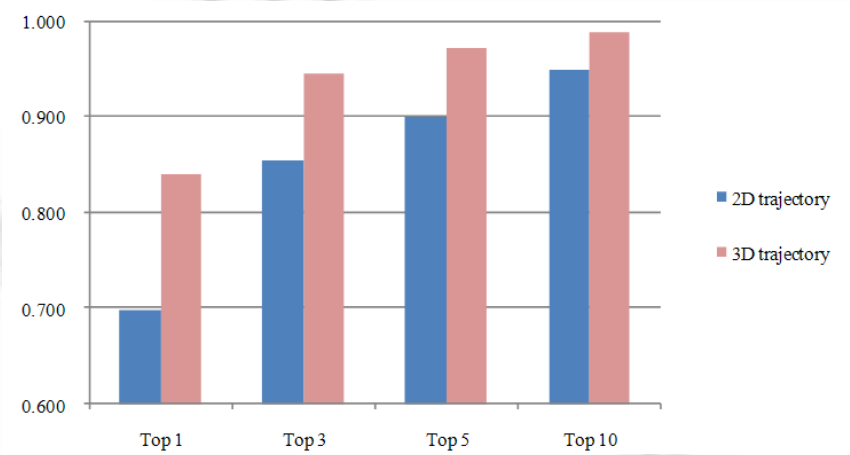


Fig.7 Recognition rate contrast of 2D and 3D trajectory

图 7 二维和三维轨迹的识别率对比

图 7 给出了分别基于二维轨迹和三维轨迹的识别结果的对比.从中可以看出,三维轨迹在手语识别中的表现显然要远胜于二维轨迹.在 Top1 上,三维轨迹的识别率大约比二维轨迹高 14%.当 Top 的数目变大时,它们之间的差距略微缩小,但依然十分显著.另一方面,我们的方法在使用了三维轨迹后的 Top1 识别率达到了 83.9%,在 Top3 上则达到了 94.5%(见表 2).这说明了我们使用三维轨迹进行手语识别的方法是有效的.

4 结 论

轨迹是手语最重要的特征之一.本文提出了一种基于频域上三维轨迹匹配的高效的手语识别方法.每条原始的轨迹都被归一化,并重采样成均匀的曲线.然后使用 DFT 来生成更稳定更有区分性的特征.轨迹的匹配被转化成两组 DFT 系数的实部、虚部和复数模匹配.最后,将根据最大的匹配得分来得出最终的识别结果.实验结果显示,3D 轨迹在手语识别中的表现比 2D 要好.此外,我们基于 3D 轨迹的手语识别方法在 239 个手语词的数据集上也表现良好.另一方面,由于采用了简便且价格低廉的 Kinect 作为数据采集的设备,相对于以往的基于数据手套等昂贵设备的方法,也更加适于推广.

上述的工作只是考虑了孤立词的识别,而我们接下去的目标是连续的(例如以句子为单位)手语识别.当输入是一个连续的句子时,其中包含大量词语间的过渡轨迹,而这些轨迹是不属于任何词汇的.根据轨迹运动的特点,可以用相对静止的关键帧作为分词的初始节点,根据初始节点进行部分曲线匹配的遍历结果进行分词.另一方面,中国手语总共有将近 5 000 个词汇,我们应当在一个更大词汇集上保证该方法的有效性.从数据的特点上看,大规模的词汇集含有大量轨迹特征相似的静态词语,单纯的轨迹特征对这样的词汇是不够的.因此,如何结合另一个手语的重要特征——手型,也将是我们下一步的研究方向.

References:

[1] Murakami K, Taguchi H. Gesture recognition using recurrent neural networks. In: Robertson SP, Olson GM, Olson JS, eds. Proc. of the Special Interest Group on Computer-Human Interaction Conf. on Human Factors in Computing Systems: Reaching Through Technology. New York: ACM, 1991. 237−242.

[2] Huang CL, Huang WY, Lien CC. Sign language recognition using 3D Hopfield neural network. In: Proc. of the Int'l Conf. on Image Processing. Washington: IEEE, 1995,2:611−614.

[3] Kim JS, Jang W, Bien Z. A dynamic gesture recognition system for the Korean signlanguage (KSL). IEEE Trans. on Systems, Man, and Cybernetics, Part B: Cybernetics, 1996,26(2):354−359.

[4] Grobel K, Assan M. Isolated sign language recognition using hidden Markov models. In: Proc. of the IEEE Int'l Conf. on Systems, Man, and Cybernetics. Orlando: IEEE, 1997,1:162−167.

[5] Starner T, Weaver J, Pentland A. Real-Time American sign language recognition using desk and wearable computer based video. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1998,20(12):1371−1375.

[6] Mitra S, Acharya T. Gesture recognition: A survey. IEEE Trans. on System, Man, and Cybernetics, Part C: Applications and reviews, 2007,37(3):311−324.

[7] Yang R, Sarkar S. Coupled grouping and matching for sign and gesture recognition. Computer Vision and Image Understanding, 2009,113(6):663−681.

[8] Shi J, Zhang M, Pan Z. A real-time bimanual 3D interaction method based on bare-hand tracking. In: Proc. of the 19th ACM Int'l Conf. on Multimedia. Scottsdale: ACM, 2011. 1073−1076.

[9] Zhu J, Wang L, Yang R, Davis JE, Pan Z. Reliability fusion of time-of-flight depth and stereo geometry for high quality depth maps. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2011,33(7):1400−1414.

[10] Ren Z, Meng J, Yuan J, Zhang Z. Robust hand gesture recognition with Kinect sensor. In: Proc. of the 19th ACM Int'l Conf. on Multimedia. Scottsdale: ACM, 2011. 759−760.

[11] Zafrulla Z, Brashear H, Starner T, Hamilton H, Presti P. American sign language recognition with the kinect. In: Proc. of the 13th Int'l Conf. on Multimodal Interfaces. Alicante: ACM, 2011. 279−286.

[12] Tong J, Zhou J, Liu L, Pan Z, Yan H. Scanning 3D full human bodies using Kinects. IEEE Trans. on Visualization and Computer Graphics, 2012,18(4):643−650.

[13] Wang J, Liu Z, Wu Y, Yuan J. Mining actionlet ensemble for action recognition with depth cameras. In: Proc. of the Conf. on Computer Vision and Pattern Recognition (CVPR). Providence: IEEE, 2012. 1290−1297.

[14] Zhang Y, Zhang S, Luo Y, Xu XD. Gesture track recognition based on Kinect depth image information and its application. Application Research of Computer, 2012,29(9):3547−3550 (in Chinese with English abstract).

[15] Deng R, Zhou LL, Ying RD. Gesture extraction and recognition research based on Kinect depth data. Application Research of Computer,2013,30(4):1263−1265 (in Chinese with English abstract).

[16] Wobbrock JO, Wilson AD, Li Y. Gestures without libraries, toolkits or training: A $1 recognizer for user interface prototypes. In: Shen C, Jacob R, Balakrishnan R, eds. Proc. of the 20th Annual ACM Symp. on User Interface Software and Technology. Newport: ACM, 2007. 159−168.

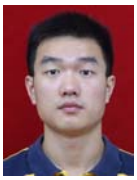
附中文参考文献:

[14] 张毅,张烁,罗元,徐晓东.基于 Kinect 深度图像信息的手势轨迹识别及应用.计算机应用研究,2012,29(9):3547−3550.

[15] 邓瑞,周玲玲,应忍冬.基于 Kinect 深度信息的手势提取与识别研究.计算机应用研究,2013,30(4):1263−1265.

林宇舜(1988－),男,广东湛江人,博士生,主要研究领域为计算机视觉,模式识别.
E-mail: yushun.lin@vipl.ict.ac.cn

柴秀娟(1978－),女,博士,副研究员,主要研究领域为计算机视觉,模式识别,人机交互.
E-mail: xiujuan.chai@vipl.ict.ac.cn



许志浩(1990－),男,硕士生,主要研究领域为计算机视觉,模式识别.
E-mail: zhihao.xu@vipl.ict.ac.cn

尹芳(1989－),女,博士生,主要研究领域为计算机视觉,模式识别
E-mail: fang.yin@vipl.ict.ac.cn

陈熙霖(1965－),男,博士,研究员,主要研究领域为图像理解,计算机视觉,模式识别,图像处理.
E-mail: xilin.chen@vipl.ict.ac.cn