

一种可处理数据缺失的视角无关手语识别方法

王 骐¹⁾ 陈熙霖²⁾ 王春立³⁾ 高 文^{2),4)}

¹⁾(哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)

²⁾(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)

³⁾(大连海事大学信息科学与技术学院 辽宁 大连 116026)

⁴⁾(北京大学信息科学与技术学院 北京 100871)

摘 要 基于虚拟立体视假设,借鉴 RANSAC 技术的思想,文中针对数据缺失(帧对之间匹配特征可能较少)情况下的视角无关手语识别问题,提出一种 Sample-Consensus 方法.其基本出发点是,同一手语不同视角下的两个样本序列之间所有的对应帧对,可以解释为由某一虚拟立体视觉系统同步捕获,因而满足同一个基础矩阵,而且此基础矩阵能够基于部分对应帧对包含的点对应关系进行估计.实验表明,提出的 Sample-Consensus 方法能够有效地应用于数据缺失情况下的视角无关手语识别.另外,这种方法也可以扩展到相近的领域,如视角无关的动作识别和刚体运动分析等.

关键词 手语识别;数据缺失;极线约束;随机抽样一致性方法

中图法分类号 TP391

DOI号: 10.3724/SP.J.1016.2009.00953

A Data-Deficiency-Tolerated Method for Viewpoint Independent Sign Language Recognition

WANG Qi¹⁾ CHEN Xi-Lin²⁾ WANG Chun-Li³⁾ GAO Wen^{2),4)}

¹⁾(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001)

²⁾(Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

³⁾(School of Information Science & Technology, Dalian Maritime University, Dalian, Liaoning 116026)

⁴⁾(School of Electronics Engineering and Computer Science, Peking University, Beijing 100871)

Abstract This paper proposes a Sample-Consensus method for viewpoint independent sign language recognition under data deficiency (matched features are possibly deficient with regard to some frame pairs). The proposed method is based on the epipolar geometry and inspired by RANSAC. The basic idea is that all corresponded frames between two sequences of the same sign can be roughly considered as captured synchronously by a virtual stereo vision system and thus they will satisfy the same fundamental matrix. In addition, the fundamental matrix can be estimated from point correspondences contained by some part of corresponding frames. Experimental results demonstrate the efficiency of the proposed method. Moreover, this Sample-Consensus method can be easily extended to some similar problems, such as viewpoint independent activity analysis and rigid-motion analysis.

Keywords sign language recognition; data deficiency; epipolar constraint; RANSAC

收稿日期:2008-03-24;最终修改稿收到日期:2008-12-18. 本课题得到国家自然科学基金(60533030,60603023,U0835005)、多媒体与智能软件技术北京市重点实验室开放课题资助. 王 骐,男,1979年生,博士研究生,主要研究方向为计算机视觉、视角无关的手语识别. 陈熙霖(通信作者),男,1965年生,博士,研究员,主要研究领域为计算机视觉、模式识别、多模式接口、数字电视技术等. E-mail: xlchen@ict.ac.cn. 王春立,女,1972年生,博士,教授,主要研究领域为模式识别、人机交互. 高 文,男,1956年生,博士,教授,IEEE Fellow,主要研究领域为计算机视觉、模式识别与图像处理、多媒体数据压缩、多模式接口、虚拟现实、视频编码与分析、手语识别与合成、人脸识别、数字图书馆等.

1 引 言

手语是聋哑人进行信息交流最自然的方式. 研究手语识别的目的, 是希望通过计算机将手语翻译成文本或语音, 为聋人和听力正常人之间的交流架起一座方便的桥梁. 此外, 手语识别还可应用于虚拟环境中控制虚拟人的运动以及在虚拟现实作为多模式接口.

关于手语识别方面的综述可以参见文献[1]. 从数据获取方式上分, 手语识别主要有基于数据手套的手语识别和基于视觉的手语识别两种. 前者由于能采集到准确的手语数据, 可以获得较高的识别精度. 但数据手套代价昂贵, 容易损坏, 使用起来也很不方便. 相比之下, 使用摄像机的基于视觉的方式更方便, 更自然.

然而, 大部分基于视觉的手语识别研究都限制在特定视角之下, 这往往是由于受限于研究者所使用的特征. Starner 和 Pentland^[2]利用 HMM 模型进行连续手语识别的研究, 所采用的特征向量共包含有 8 个分量, 分别为每只手的 X、Y 坐标, 每只手外接椭圆的主轴方向和外接椭圆的偏心率. 这些分量均依赖于特定的视角. Bauer 和 Hienz^[3]也进行连续手语识别的研究, 为方便特征提取, 他们采用了颜色手套, 所提取的特征包括每只手相对于身体中心线的 X 坐标和相对于肩膀的 Y 坐标、两只手重心之间的距离、各个颜色区域重心的相对距离以及各颜色区域的尺寸等. 这些特征分量也都依赖于相应的视角. Bowden 等人^[4]提出了一种语义特征, 用于手势识别. 这种语义特征本身对摄像机的摆放能表现出一定程度的无关性, 然而, Bowden 等人所采用的提取这种语义特征的方法却对摄像机的摆放有着一定程度的依赖. 在基于视觉的中国手语识别研究上, 所使用的特征大多数也与视角有着密切的关系, 如李勇等^[5]从颜色手套中提取的由颜色重心距离构造而成的手形特征多边形以及张良国等^[6]所使用的颜色手套特征等. 上面的这些研究都使用视角相关的特征, 而采用的识别技术又不能处理视角的变化, 因此这些方法都只能工作在特定视角之下.

限制在一个特定视角之下意味着手语者只能在特定的空间以特定的朝向执行手语, 这严重地制约着手语者的自由. 因此, 有必要研究视角无关的手语识别, 以解除对摄像机捕获视角的限定, 从而方便用户的使用.

基于 3 个正交放置的摄像机, Vogler 和 Metaxas^[7]通过计算机视觉方法提取出手语者手的三维运动参数, 并利用 HMM (Hidden Markov Models) 进行识别. 针对 53 个美国手语孤立词, 识别率为 89.9%. 因为三维特征独立于摄像机的捕获视角, 这种方式可以获得视角无关的手语识别. 然而, 由于使用多个摄像机在现实应用中是个较强的限制, 同时也由于估计三维参数本身对噪声很敏感, 基于多个摄像机提取三维参数的视角无关手语识别方式难于实际运用.

针对静态手指语, Wu 和 Huang^[8]提出一种基于表观的学习方法, 获得了一定程度的视角无关性. 针对 14 个手型, 使用上万的样本, 识别率为 92.4%. 然而, 针对动态手语, 这种基于表观的学习方法不能应用. 原因在于动态手语中, 丰富的手型变换, 同时受视角变化的影响, 使得动态手语的表观变化非常复杂. 在这样的情况下使用基于表观的学习方法, 一来模型不易建立, 二来所需样本数量的巨大也会给样本收集和模型训练带来很大困难.

不同于上面两个工作, 作者主要研究利用极线几何约束实现单摄像机前提下视角无关的动态手语识别. 与此相关的一个工作是 Rao 等人^[9]同样基于极线几何技术所进行的视角无关的活动匹配研究, 他们通过使用第 9 维奇异值约束较好地解决了同一活动不同视角下的活动序列之间的匹配问题. 然而, 对于手语识别这类精细差别较多的多类别问题, 仅仅使用第 9 维奇异值约束不能取得满意的结果.

在前期工作^[10]中, 作者提出一种在对齐观测序列和模板序列后、通过对应帧验证基础矩阵唯一性的视角无关手语识别方法, 并引入证据理论, 在合理分析对齐后两个序列之间的相似成分和相异成分而形成的相似度证据和相异度证据的基础上, 将基础矩阵唯一性验证转化为证据理论框架下的一个判别问题, 以提高对匹配误差的鲁棒性. 实验表明了这种方法的有效性.

前期工作由于需要在每个匹配帧对之间估计一个基础矩阵, 因此要求每个匹配帧对包含至少 8 个特征点对. 在实际应用中, 受视角变化所引起的遮挡或者特征点漏检等因素的影响, 很可能某些帧对之间的对应特征点对数目会少于 8. 在这些情况下, 前期工作提出的验证方法不能工作.

与前期工作^[10]相比, 本文重点研究数据缺失情况下, 即由于单个帧对之间特征点对数目较少而使得算法无法工作情况下的视角无关手语识别问题.

注意到,同一手语不同视角下的两个样本序列之间所有的对应帧对均满足同一个基础矩阵^①,且该基础矩阵能够基于部分对应帧对包含的点对应进行估计.从这一点出发,借鉴 RANSAC 技术^[11]的思想,本文提出一种 Sample-Consensus 方法,用于数据缺失情况下的视角无关手语识别.由于是基于多个帧对进行基础矩阵的估计,提出的方法能够在很大程度上避免前期工作中基于单个帧对进行基础矩阵估计所可能遇到的数据缺失情况.实验表明了所提出方法的有效性.同时,这种 Sample-Consensus 方法能够容易地扩展到相近的领域,比如视角无关的动作识别和刚体运动分析等.

本文第 2 节描述提出的 Sample-Consensus 方法;第 3 节给出实验结果;第 4 节总结全文并给出将来的方向.

2 提出的方法

如前所述,由于注意到同一手语不同视角下的两个样本序列所有的对应帧对均满足同一个基础矩阵,且该基础矩阵能够基于部分对应帧对包含的点对应进行估计,因而提出一种 Sample-Consensus 方法,用于数据缺失情况下的视角无关手语识别.其基本框架如图 1 所示.

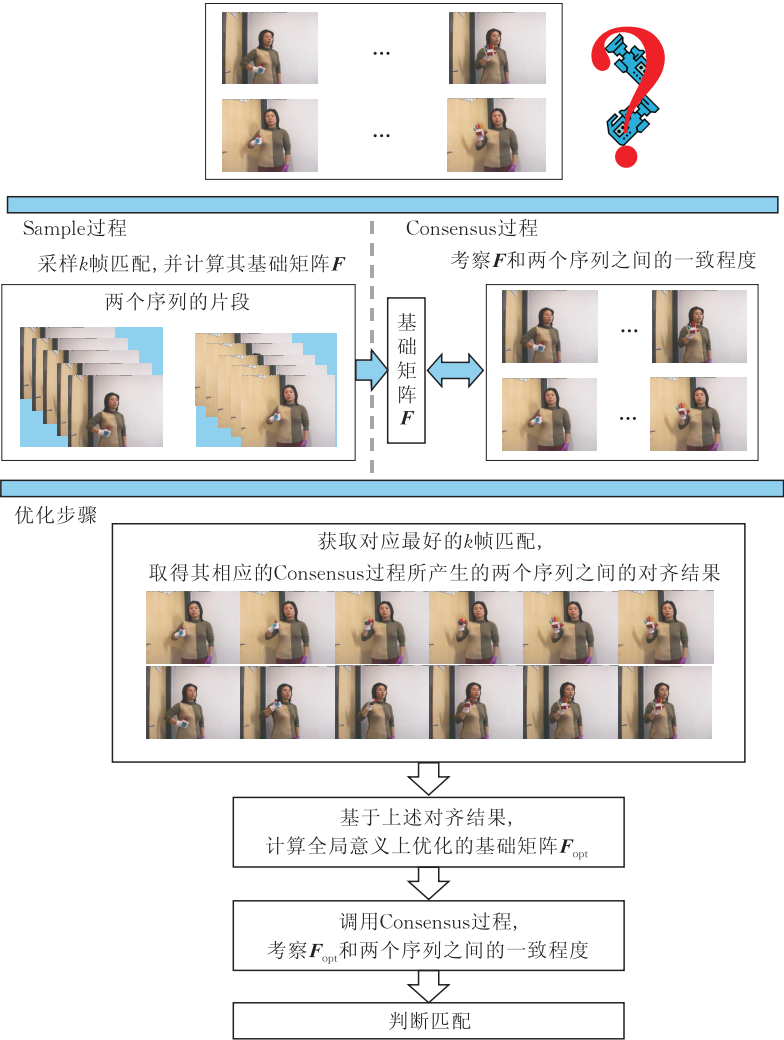


图 1 Sample-Consensus 方法基本框架

提出的 Sample-Consensus 方法包含两个核心过程——Sample 过程和 Consensus 过程,并按如下方式工作:Sample 过程负责采样观测序列和当前模板序列片段之间(如前 m_1 帧和后 m_2 帧中)所有可能的 k 帧匹配(不失一般性,用帧序号记之为

① 在本文中,如果说基础矩阵 F 与点对应 (p_0, p_1) 一致,或者说点对应 (p_0, p_1) 与基础矩阵 F 一致,是说点对应 (p_0, p_1) 能够构成在以 F 为基础矩阵的某个立体视觉系统下的对应图像点对;如果说某一帧对满足基础矩阵 F ,或者说基础矩阵 F 满足某一帧对,是指此帧对包含的所有感兴趣特征点对与基础矩阵 F 一致.

$((i_1, j_1), (i_2, j_2), \dots, (i_k, j_k)))$, 并针对每一可能的 k 帧匹配, 在假定其正确对应的前提下, 从所包含的点对应中估计出相应的“基础矩阵”; Consensus 过程负责度量一个基础矩阵和两个序列之间的一致程度; 通过这两个过程, 获得所有可能性中最接近正确对应的 k 帧匹配; 并在此基础上, 通过一个优化步骤, 给出观测序列和当前模板序列之间的一个近似度度量; 最后, 通过最近邻法则进行识别. 由于是基于多个帧对进行基础矩阵的估计, 提出的 Sample-Consensus 方法能够在很大程度上避免前期工作中基于单个帧对进行基础矩阵估计所可能遇到的数据缺失情况.

接下来的部分首先讨论特征点定位和匹配, 然



图 2 颜色手套

由于本文重心是测试提出的 Sample-Consensus 方法针对视角变化和数据缺失情况的执行性能, 本文不对特征点的自动检测问题特别关注. 关于特征点的定位和匹配, 本文直接采用人工指定的方式实现. 另外还需指出的是, 本文工作只涉及单主手词汇.

2.2 Sample-Consensus 方法

提出的 Sample-Consensus 方法在假定观测序列和当前模板序列代表同一手语的前提下, 负责给出它们之间的一个近似度度量, 而无论其视角变化和某些帧对之间是否存在数据缺失. 其核心过程包括一个 Sample 过程和一个 Consensus 过程, 并按下述方式工作.

Sample 过程. RANSAC 技术的核心思想是, 通过采样, 发现一组误差较小的数据, 使得从其中估计

后详细介绍提出的 Sample-Consensus 方法.

2.1 特征点定位和匹配

使用极线几何约束, 首先需要解决帧图像上特征点定位和帧对之间特征点匹配的问题. 为便于特征点精确定位、帧对之间特征点匹配以及更精细地刻画手的姿态, 本文工作采用一对颜色手套, 其设计如下: 在手语中, 虽然两只手都有可能被用到, 但两只手所起作用并不一样, 其中一只手起主要作用, 称为主手, 通常是右手, 另一只手起辅助作用, 称为副手, 通常是左手; 考虑到绝大部分的手语信息主要由主手的运动来表达, 主手手套用 7 种颜色, 分别着色于 5 个手指、手心 and 手背, 副手手套用另外一种颜色统一着色. 设计出的颜色手套如图 2 所示.

出的模型能够最大限度地满足所有数据. 类似的, 这里的 Sample 过程负责采样到一个正确对应的 k 帧匹配, 以使得从其中估计出的基础矩阵能够最大限度地满足两个序列之间所有的对应帧对.

观测序列和模板序列代表同一手语的假定, 预示着在两段序列的前 m 帧中至少存在一个正确对应的 k 帧匹配 ($k < m$, 并合理取值). 因此, 采样前 m 帧中所有可能的 k 帧匹配能够确保采样到一个正确对应的 k 帧匹配.

然而, 仅仅保证采样到一个正确对应的 k 帧匹配是不够的. 因为如果这个 k 帧匹配提供的特征点对之间不相互独立, 则可能造成事实上的数据缺失, 而无法保证基础矩阵估计的顺利进行. 从两段序列的前 m 个连续帧中采样到的 k 帧匹配, 由于运动上的连续性, 很有可能提供的特征点对之间并不相互

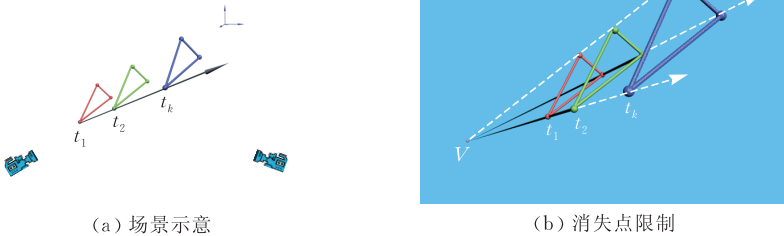


图 3 一个正确对应、能提供 $3 \times k$ 个特征点对、但仅有 5 个独立点对的 k 帧匹配例子

独立. 考虑图 3(a)所示的场景, 空间中的 3 个点进行着一致的直线运动. 在左右两个视角之间, 由此场景形成的 k 帧匹配所能提供的特征点对为 $3 \times k$ 个, 然而其中相互独立的特征点对却仅仅只有 5 个. 原因是, 一致的直线运动在相同的成像平面上对应着唯一的消失点^[12], 而这个消失点可以由对应于两个物理特征点的 4 个图像点唯一确定, 这一点可直接从图 3(b)中看出.

在实际执行中, 为保证能够提供足够的且相互独立的特征点对, Sample 过程从两段序列的前 m_1 帧中采样一个 k_1 帧匹配, 从后 m_2 帧中采样一个 k_2 帧匹配, 然后合成一个 k 帧匹配 ($k_1 + k_2 = k$). 理由为, 对于图 3(a)所示场景形成的任意 k 帧匹配, 如果其中某一帧对偏离原来的直线运动, 都将确保提供 8 个独立的特征点对 ($5 + 3$), 而 8 个独立的特征点对对基础矩阵的估计是充分的. 既然绝大多数手语词首尾两段包含有不一致的运动, 提出的解决方案将保证采样到的 k 帧匹配提供数量足够的且相互独立的特征点对, 从而确保基础矩阵估计工作的顺利进行.

对每个采样的 k 帧匹配, Sample 过程也负责在假定其正确对应的前提下, 基于 8 点算法和使用 RANSAC 技术从包含的点对中估计相应的基础矩阵.

Consensus 过程. Consensus 过程负责度量一个基础矩阵 F 与一个序列对 (即观测序列和当前的模板序列) 之间的一致程度. 从极线几何约束可知, 某个基础矩阵和某个序列对之间相互一致, 当且仅当序列对中的两个序列匹配, 同时这个基础矩阵与这两个序列之间所有的对应帧对一致. 因此, 本文提出如下度量方法: 首先, 在基础矩阵 F 的作用下, 通过动态时间规整 (Dynamic Time Warping, DTW) 对齐两个序列; 然后, 度量 F 与对齐后两个序列之间所有匹配帧对的一致程度.

动态时间规整中的一个基本步骤是评价帧对之间的匹配程度, 因此数据缺失是需要考虑的一个问题, 因为在数据缺失情况下, 某些帧对之间可能只有很少量匹配的特征点对, 甚至可能完全没有匹配的特征点对. 如何评价数据缺失情况下帧对之间的匹配程度, 是需要考虑的一个关键问题. 据前面介绍, 两个序列之间所有匹配帧对满足同一个基础矩阵. 因而, 单个帧对是否匹配等价于此帧对包含的点对是否与基础矩阵一致. 基于这一点, 提出如下解决方法.

由极线几何约束可知, 一个点对 (p_0, p_1) , 如果它们在由基础矩阵 F 决定的立体视觉系统中是对应图像点对, 则它们与基础矩阵 F 将满足如下的关系式^[12]:

$$p_0^T F p_1 = 0 \quad (1)$$

式(1)的几何意义可以解释为: 在基础矩阵 F 决定的立体视觉系统中 (见图 4), 对于左边视角下的图像点 p_0 , 它在右边视角下所有可能的对应点 (包括 p_1) 都将被限制在直线 $F^T p_0$ 上; 反之, 对右边视角下的图像点 p_1 , 它在左边视角下所有可能的对应点 (包括 p_0) 都将被限制在直线 $F p_1$ 上. 这两个解释可以形式化为

$$\begin{cases} d^2(p_0, F p_1) = 0 \\ d^2(p_1, F^T p_0) = 0 \end{cases} \quad (2)$$

其中 d 代表空间中点到直线的距离.

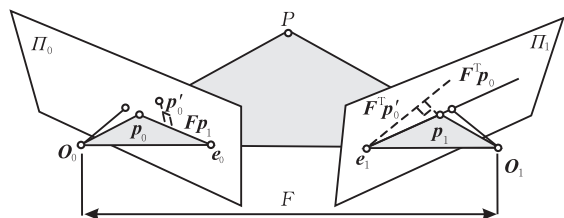


图 4 极线几何约束

定义:

$$E(F, (p_0^*, p_1)) = \frac{1}{2} (d^2(p_0^*, F p_1) + d^2(p_1, F^T p_0^*)) \quad (3)$$

此定义显示, 当 p_0^* 与 p_0 重合时, $E(F, (p_0^*, p_1))$ 等于 0; 当 p_0^* 远离 p_0 时 (参考图 4 中 p'_0), $E(F, (p_0^*, p_1))$ 将逐渐变大. 这意味着 $E(F, *)$ 的大小反映着点对在基础矩阵为 F 的立体视觉系统中的对应程度, 也即反映着点对与基础矩阵 F 的一致程度.

对 N 个点对 (记为点对集 A) 和基础矩阵 F , 定义平均匹配程度:

$$\begin{aligned} C(F, A) &= \frac{1}{\frac{1}{N} \sum_{i=1}^N E(F, (p_{0i}, p_{1i})) + \alpha} \\ &= \frac{1}{\frac{1}{2N} \sum_{i=1}^N (d^2(p_{0i}, F p_{1i}) + d^2(p_{1i}, F^T p_{0i})) + \alpha} \end{aligned} \quad (4)$$

其中设置常数 $\alpha = 1$ 的目的是为了归一化匹配程度的值.

由于 $E(F, *)$ 反映点对与基础矩阵 F 的一致程度, 平均匹配程度 $C(F, A)$ 能够反映 A 中的 N 个

点对和基础矩阵 F 之间的一致程度. 由式(3)和式(4)可知, 如果 $C(F, A)$ 越大, 则意味着 A 中的 N 个点对和基础矩阵 F 之间的一致程度越高.

既然评价单个帧对的匹配程度, 或者评价 DTW 网格中的某条路径的匹配程度, 都等同于评价一组点对应与基础矩阵 F 的一致程度, 基于 $C(F, *)$, 使用 DTW 即可实现数据缺失情况下观测序列和模板序列之间的对齐. 随之而获得的 DTW 距离, 记为 C_F , 即可被用来度量基础矩阵 F 与由观测序列和当前模板序列构成的序列对之间的一致程度.

基于上面描述的 Sample 过程和 Consensus 过程, 提出的 Sample-Consensus 方法通过如下步骤来度量观测序列和模板序列之间的匹配程度:

1. 利用 Sample 过程采样两个序列中的部分片段之间所有可能的 k 帧匹配并计算相应的基础矩阵 F ;
2. 利用 Consensus 过程, 对 Sample 过程采样出的所有 k 帧匹配相应的基础矩阵 F 进行测试, 以发现其中最接近正确对应的 k 帧匹配(与之相应的基础矩阵与观测序列和模板序列构成的序列对之间具有最高的一致程度);
3. 取出相似度最高的 k 帧匹配相应的 Consensus 过程所获得的两个序列之间的对齐结果, 并在此基础上, 计算出一个优化的基础矩阵 F_{opt} ; 然后, 调用 Consensus 过程, 计算获得的 F_{opt} 与观测序列和模板序列构成的序列对之间的一致程度, 记获得的结果为 $C_{F_{\text{opt}}}$; 利用 $C_{F_{\text{opt}}}$ 量度观测序列和模板序列之间的近似程度.

从步骤 3 中可以看出, F_{opt} 代表全局意义上优化的基础矩阵, 因此, 基于 F_{opt} 所计算出的 $C_{F_{\text{opt}}}$ 将能够有效地反映出观测序列和模板序列之间的近似程度.

2.3 识 别

从 2.2 节的描述可知, 从两个手语序列中计算出的 $C_{F_{\text{opt}}}$ 越大, 表明两个手语序列之间的近似程度越高. 将 $C_{F_{\text{opt}}}$ 作为近似度度量, 通过最近邻法则, 即可实现数据缺失情况下视角无关的手语识别. 识别过程形式如下:

$$\begin{aligned} R(S_0) &= \underset{\substack{S_k \in \text{the template set} \\ k=1, 2, \dots, K}}{\operatorname{argmax}} \quad C_{F_{\text{opt}}}^{S_0, S_k} \\ &= \underset{\substack{S_k \in \text{the template set} \\ k=1, 2, \dots, K}}{\operatorname{argmax}} \quad 1 / \left(\frac{1}{2N} \sum_{i=1}^N \left(d^2(p_{0i}, F_{\text{opt}} p_{1i}) + \right. \right. \\ &\quad \left. \left. d^2(p_{1i}, F_{\text{opt}}^T p_{0i}) \right) + \alpha \right) \end{aligned} \tag{5}$$

其中 K 代表模板的数量, 在手语识别中为所有词汇的数量, F_{opt} 代表利用提出的 Sample-Consensus 方法所计算出的观测序列 S_0 和模板序列 S_k 之间全局意义上优化的基础矩阵, N 代表 S_0 和 S_k 之间、在基

础矩阵 F_{opt} 作用之下所有的对应帧对包含的特征点对数目.

3 实验结果与分析

为评价 Sample-Consensus 方法的有效性, 本文在包含有 125 个中国手语词汇的中等词汇集上对其进行测试. 实验数据集包含 3 个手语样本集, 分别称为样本集 1、样本集 2 和样本集 3, 其中样本集 1 从正面视角进行采集, 样本集 2 也从正面视角进行采集但摄像机摆放位置与样本集 1 不同, 样本集 3 从一个大约 30° 的侧面视角进行采集. 所有的手语数据均由熟练的手语者执行, 并在实验室环境中通过 USB 彩色摄像机收集而得.

Sample-Consensus 方法能够处理视角的变化, 并能处理数据部分缺失情况. 本节将分别测试 Sample-Consensus 方法在这两种情况下的执行性能并与有关的一些方法进行比较. 在 Sample-Consensus 方法所有执行中其参数设置固定, 即 m_1 、 k_1 、 m_2 、 k_2 分别取 5、2、3、1, 也就是从观测序列和模板序列的前 5 帧中采样一个 2 帧匹配, 从后 3 帧中采样一个 1 帧匹配, 合成一个 3 帧匹配, 然后基于此 3 帧匹配进行基础矩阵的估计.

3.1 对视角变化的有效性

本小节测试 Sample-Consensus 方法对于视角变化的执行性能, 过程是轮流采用 3 个样本集中的一个样本集作为模板集, 另两个样本集分别作为观测集, 共形成六组模板-观测对, 用于测试提出方法的识别性能. 另外, 作为对比, 同样也测试了采用 Rao 等人^[9]提出的第 9 维奇异值约束以及前期工作提出的验证方法^[10]在同样的数据集上的识别性能, 结果如图 5 所示.

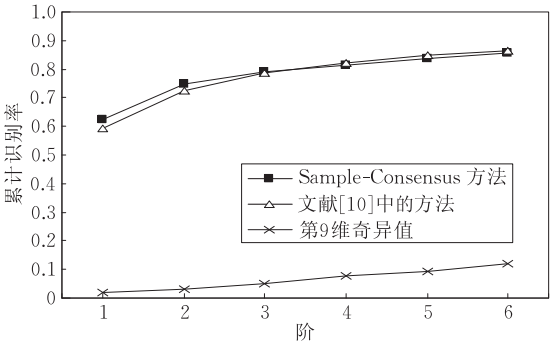


图 5 Sample-Consensus 方法和相关工作针对视角变化的执行性能比较

图 5 使用 CMS(Cumulative Match Scores) 曲线给出了各种方法平均识别性能的一个直观示意。

从图 5 可以看出,使用 Rao 等用于视角无关对齐的第九维奇异值约束,获得的首选识别率仅为 1.9%,获得的 Top 3 识别率也仅为 5.1%,这表明第 9 维奇异值不能有效地应用于视角无关的手语识别.其原因在于,正如 Rao 等在文献[9]中所声称的那样,第 9 维奇异值是一种对噪声非常敏感的近似度量。

相比之下,作者前期工作^[10]提出的基于证据理论的基础矩阵唯一性验证方法分别获得了 59.3% 的首选识别率和 78.8% 的 Top 3 识别率,本文提出的 Sample-Consensus 方法分别获得了 62.5% 的首选识别率和 79.1% 的 Top 3 识别率,这充分地表明了这两种方法对于视角无关手语识别的有效性。

3.2 对数据缺失的稳定性

本小节测试 Sample-Consensus 方法针对数据缺失的性能。

首先基于样本集 2 通过随机移除某些特征点生成两个具有不同数据缺失程度的观测集,数据缺失集 2-1 和数据缺失集 2-2. 数据缺失集 2-1 中,每一观测序列的每一帧包含 4~7 个特征点. 数据缺失集 2-2 相对于数据缺失集 1,具有更严重的数据缺失程度. 数据缺失集 2-2 中,每一观测序列有 2/3 的帧每一帧包含 4~5 个特征点,剩下的 1/3 的帧每一帧仅包含 0~3 个特征点. 在生成的这两个数据缺失观测

集上,前期工作中的验证方法^[10]不能工作,原因是单个帧对之间包含至少 8 个特征点对的前提条件不能得到满足。

测试过程是轮流采用数据缺失集 2-1 和数据缺失集 2-2 作为观测集,样本集 1 和样本集 3 作为模板集,共形成 4 组模板-观测组合,用于测试提出方法在数据缺失情况下的执行性能. 测试结果如表 1 和图 6 所示。

表 1 Sample-Consensus 方法对数据缺失测试结果						
组合	Top 1/%	Top 2/%	Top 3/%	Top 4/%	Top 5/%	Top 6/%
(1,2)	62.4	73.6	76.8	79.2	83.2	86.4
(1,2-1)	50.4	66.4	72.8	76.0	80.0	81.6
(1,2-2)	40.8	52.8	60.8	71.2	73.6	76.8
(3,2)	60.8	74.4	78.4	80.0	84.0	87.2
(3,2-1)	58.4	74.4	76.0	78.4	80.8	84.0
(3,2-2)	44.8	60.8	69.6	76.8	80.0	82.4

注:组合(a,b)的意思是:样本集 a 作为模板集,而样本集 b 作为观测集。

表 1 和图 6 表明:在以样本集 1 作为模板集时, Sample-Consensus 方法对数据缺失集 2-1 的首选识别率为 50.4%,落后于对样本集 2 的首选识别率(62.4%)12 个百分点,但其 Top 3 识别率达到 72.8%,仅落后于对样本集 2 的 Top 3 识别率 4 个百分点;而在以样本集 3 作为模板集时, Sample-Consensus 方法对数据缺失集 2-1 获得的识别结果几乎与针对样本集 2 持平. 这一结果表明, Sample-Consensus 方法能够有效地处理一定程度的数据缺失情况。

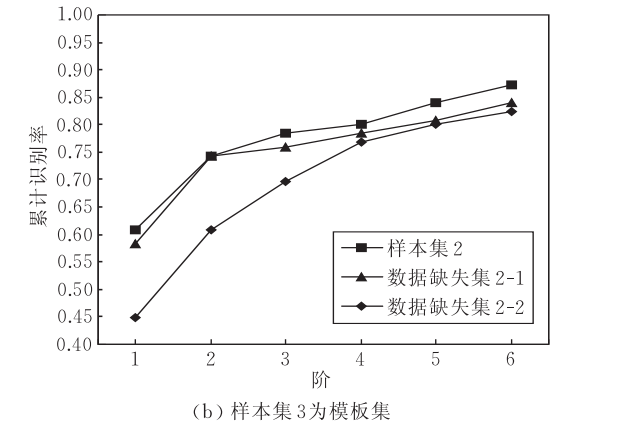
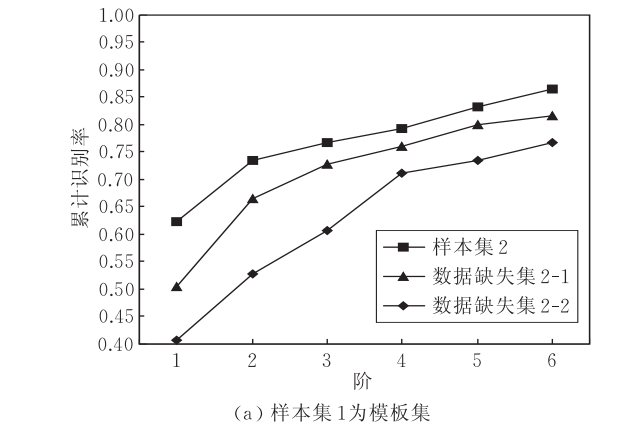


图 6 Sample-Consensus 方法针对数据缺失的执行性能

表 1 和图 6 也表明:无论是以样本集 1 还是样本集 3 作为模板集, Sample-Consensus 方法针对数据缺失集 2-2 的首选识别率均较针对数据缺失集 2-1 有大的下降,分别下降 9.6 和 13.6 个百分点;但 Sample-Consensus 方法针对数据缺失集 2-2 所获

得的 Top 4 识别率却与针对数据缺失集 2-1 相差无几. 这反映出 Sample-Consensus 方法针对更为严重的数据缺失也具有一定程度的有效性. Sample-Consensus 方法针对数据缺失集 2-2 首选识别率之所以有大的下降,原因主要有两个方面:(1)严重的

数据缺失情况下,基础矩阵 F_{opt} 不能得到很好的估计;(2)严重的数据缺失使得无法获得充分的信息以鉴别相近的手语词。

4 结 论

在前期工作^[10]的基础上,本文对利用极线和基础矩阵约束实现视角无关的手语识别进行了进一步的研究。提出的 Sample-Consensus 方法不但可以处理视角的变化,而且能够工作于数据缺失情况之下。实验表明了所提出方法的有效性。同时,这种 Sample-Consensus 方法也可以扩展到相近的领域,比如视角无关的动作识别和刚体运动分析等。实际上,凡是涉及到两个不同视角之间点对应的分析,提出的 Sample-Consensus 方法都将适用。

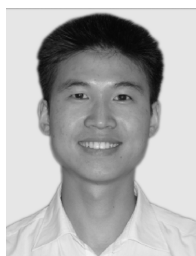
将来的研究工作涉及两个方面:(1)研究视角无关的连续手语识别;(2)研究整合多个来自于不同视角的模板的可能性。

参 考 文 献

- [1] Ong S C W, Ranganath S. Automatic sign language analysis: A survey and the future beyond lexical meaning. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 2005, 27(6): 873-891
- [2] Starner T, Pentland A. Visual recognition of American sign language using hidden Markov models//*Proceedings of the International Conference on Automatic Face and Gesture Recognition*. Zurich, Switzerland, 1995: 189-194
- [3] Bauer B, Hienz H. Relevant features for video-based continuous sign language recognition//*Proceedings of the International Conference on Automatic Face and Gesture Recognition*. Grenoble, France, 2000: 440-445
- [4] Bowden R, Windridge D, Kadir T, Zisserman A, Brady M.

A linguistic feature vector for the visual interpretation of sign language//*Proceedings of the European Conference on Computer Vision*. Prague, Czech Republic, 2004: 390-401

- [5] Li Yong, Gao Wen, Yao Hong-Xun. Chinese sign language finger alphabet recognition based on color gloves. *Computer Engineering and Applications*, 2002, 38(17): 55-58(in Chinese)
(李勇, 高文, 姚鸿勋. 基于颜色手套的中国手指字母的动态识别. *计算机工程与应用*, 2002, 38(17): 55-58)
- [6] Zhang Liang-Guo, Gao Wen, Chen Xi-Lin, Chen Yi-Qiang, Wang Chun-Li. A medium vocabulary visual recognition system for Chinese sign language. *Journal of Computer Research and Development*, 2006, 43(3): 476-482(in Chinese)
(张良国, 高文, 陈熙霖, 陈益强, 王春立. 面向中等词汇量的中国手语视觉识别系统. *计算机研究与发展*, 2006, 43(3): 476-482)
- [7] Vogler C, Metaxas D. ASL recognition based on a coupling between hmms and 3D motion analysis//*Proceedings of the IEEE International Conference on Computer Vision*. Bombay, India, 1998: 363-369
- [8] Wu Y, Huang T S. View-independent recognition of hand postures//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Hilton Head, SC, USA, 2000, 2: 88-94
- [9] Rao C, Gritai A, Shah M, Syeda-Mahmood T. View invariant alignment and matching of video sequences//*Proceedings of the IEEE International Conference on Computer Vision*. Nice, France, 2003: 939-945
- [10] Wang Q, Chen X, Zhang L, Wang C, Gao W. Viewpoint invariant sign language recognition. *Computer Vision and Image Understanding*, 2007, 108(1-2): 87-97
- [11] Fischler M A, Bolles R C. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 1981, 24(6): 381-395
- [12] Hartley R, Zisserman A. *Multiple View Geometry in Computer Vision*. 2nd Edition. New York, NY, USA: Cambridge University Press, 2003



WANG Qi, born in 1979, Ph. D. candidate. His research interests include computer vision and viewpoint free sign language recognition.

WANG Chun-Li, born in 1972, Ph. D., professor. Her research interests include pattern recognition and human-computer interaction.

GAO Wen, born in 1956, Ph. D., professor, IEEE Fellow. His research interests include computer vision, pattern recognition and image processing, image and video compression, multimodal interface, virtual reality, sign language recognition and synthesis, face recognition, and digital library.

CHEN Xi-Lin, born in 1965, Ph. D., professor. His research interests include computer vision, pattern recognition, multimodal interface, and digital video broadcasting.

Background

Sign language recognition aims to translate sign language to text or speech, so as to bridge the communication between the deaf and the hearing people and help the deaf or hard-of-hearing better integrate into the society.

According to data collection of sign language, sign language recognition is generally divided into two major categories: Dataglove-based sign language recognition and vision-based sign language recognition. Vision-based sign language recognition attracts more attention of the researchers for it is more convenient for users than Dataglove-based sign language recognition. However, the state of art of vision-based sign language recognition is far from real application due to the difficulty of extracting efficient sign language features from video or image.

One of great challenges in vision-based sign language recognition is to achieve view independent sign language recognition because view variance brings feature variance. Most of the current methods in vision-based sign language recognition require a specific view of the signers, generally the frontal view, so as to guarantee similar features between the training samples and the test samples. The constraint of a specific view means that the signers can only perform with

specific location and orientation, and limits the freedom of the signer.

The authors aim to achieve viewpoint independent sign language recognition within a certain scope with only one camera, so as to remove the restriction of the specific view and provide convenience for users. In the previous works, the authors based on the epiplor geometry and proposed the efficient method of verifying the uniqueness of fundamental matrices for viewpoint free sign language recognition. However, the method of verifying the uniqueness requires at least 8 point correspondences between one frame pair and will fail under data deficiency. So the authors propose the Sample-Consensus method in this paper for efficient viewpoint free sign language recognition under data deficiency.

This research is sponsored by the National Natural Science Foundation of China under contract Nos.60533030, 60603023 and U0835005 and partially sponsored by open project of Beijing Multimedia and Intelligent Software Key laboratory in Beijing University of Technology. One purpose of these projects is to solve the key problems in sign language recognition.