软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2016,27(Suppl.(2)):58−63

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

动态手势检测与分类*

王汉杰, 柴秀娟, 陈熙霖

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

通讯作者: 柴秀娟, E-mail: chaixiujuan@ict.ac.cn

摘 要: 提出一种对视频流中的连续手势进行检测和分类的方法.检测的目的是找到这些手势的开始帧和结束帧.提出的融合音频和视觉信息的检测方法确保了检测结果的鲁棒性和正确率.对于检测到的手势,提出一种通过在 Grassmann 流形下精确度量其协方差矩阵距离的分类方法以有效区分不同类的手势.方法在 ChaLearn Multimodal Gesture dataset 2013 上进行测试,取得了很高的识别率,Recall 和 Precision 均达到 93%以上.

关键词: 手势检测与分类;协方差矩阵;Grassmann 流形

中文引用格式: 王汉杰,柴秀娟,陈熙霖.动态手势检测与分类.软件学报,2016,27(Suppl.(2)):58−63. http://www.jos.org.cn/1000-9825/16019.htm

英文引用格式: Wang HJ, Chai XJ, Chen XL. Dynamic gesture detection and classification. Ruan Jian Xue Bao/Journal of Software, 2016,27(Suppl.(2)):58−63 (in Chinese). http://www.jos.org.cn/1000-9825/16019.htm

Dynamic Gesture Detection and Classification

WANG Han-Jie, CHAI Xiu-Juan, CHEN Xi-Lin

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

Abstract: This paper proposes a framework of gesture detection and classification in continuous sequence data. The goal of detection is to determine the start and end frame of a gesture in the continuous sequence. The detection method using multi-modal features ensures the robustness and high accuracy. To classify the detected gestures represented by covariance matrices, a distance measurement on Grassmann manifold is presented to strengthen the discriminative power. The framework is evaluated on ChaLearn Multimodal Gesture dataset 2013 and achieves high accuracy. Both Recall and Precision are higher than 93%.

Key words: gesture detection and classification; covariance matrices; Grassmann manifold

动态手势(dynamic gesture)是日常生活中人与人之间重要的沟通方式,也是一种具有表现力的体态语言.通常意义上来讲,动态手势是倾向于手部的运动,特指手部连同身体的个别部分特别是上肢的行为.其难点主要体现在连续视频流中的手势检测以及基于视觉或者其他特征的手势描述和分类[1,2].

手势检测的主要任务是确定视频流中各手势的开始和结束.Elmezain 等人[3]利用隐马尔可夫模型(hidden Markov models,简称 HMMs),通过预先定义跳转模式的方法来实时分割视频流中的 10 类手势.条件随机场(conditional random field,简称 CRF)也被应用于连续视频流中手势的检测[4].随着深度传感器的广泛应用,包括彩色、深度等特征逐渐被应用于连续视频流上的手势检测.Wu 等人[5]提取 Kinect 采集的语音、深度和彩色图像在内的特征来确定特定手势在连续视频流中的开始和结束.本文的方法利用的信息有深度、彩色图像、骨架和语音,提出了一种融合音频和视觉信息的鲁棒手势检测方法.

* 基金项目: 国家自然科学基金(61472398); 中国航天医学工程预先研究项目(2013SY54A1303)

Foundation item: National Natural Science Foundation of China (61472398); Advanced Space Medico-Engineering Research Project of China (2013SY54A1303)

收稿时间: 2015-06-01; 采用时间: 2016-01-05

在手势特征描述和分类方面,Hadfield 等人[6]通过提取局部二值模式(local binary pattern,简称 LBP)特征[7]来识别剪刀、石头、布这 3 个动态手势,达到了 95%以上的识别率.Liwicki 等人[8]则提取方向梯度直方图(histograms of oriented gradients,简称 HOG)特征来描述手型.除了手型以外,上肢的骨架点也是识别手势的重要特征.Wang 等人[9]提出了用若干骨架点的三维相对位置的变化来描述人体的动作.为了更加稳定地识别手势,本文不仅提取每一帧骨架的空间点对距离特征和手型图像的 HOG 特征作为视觉特征,还提取语音的梅尔倒频谱参数(mel-frequency cepstral coefficient,简称 MFCC)作为音频特征.协方差矩阵(covariance matrix)被用来描述手势,充分考虑所有帧之间和所有特征维度之间的相关性.协方差矩阵之间的距离通常定义在黎曼流形上,在具体计算时,黎曼流形上的距离一般通过一个近似映射投影到欧式空间进行计算.这样的近似不可避免地引入误差.因此,本文构造一种更加简洁的表示,并在 Grassmann 流形上定义两个子空间之间的距离.

本文第 1 节介绍在连续视频上的手势检测方法.第 2 节介绍手势在 Grassmann 流形上的建模和分类.第 3 节是相关的实验.最后是结论.

1 视频流中的手势检测

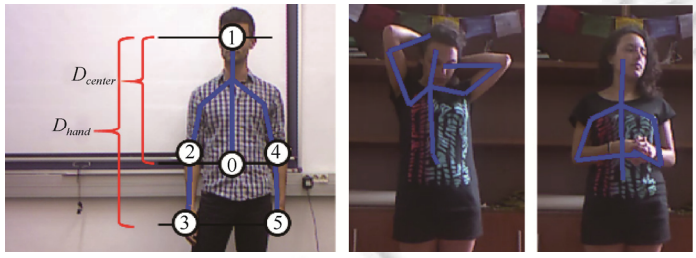
手势检测特指在连续的视频流中确定手势的开始和结束的过程.本节将分别介绍基于不同特征的手势的检测,以及最后的检测结果融合.

1.1 基于视觉的手势检测

基于视觉的手势检测是根据人体的骨架位置[10]来判断手势者在该帧中是否处于休息姿态(双手下垂,放在身体两侧).用 D_{center} 表示头部到腰部的距离,如图 1(a)所示.用 D_{hand} 表示头部到双手中最高位置手的距离.这两个距离的比值用 r 表示:

$$r = D_{hand} / D_{center} \tag{1}$$

如果 r 大于阈值 $thre$,则当前帧被判定为手势帧;反之,被判定为休息姿态帧.但是,手势者的一些习惯性小动作或者非默认的休息姿态都会被错误地判别为有意义的手势.如图 1(b)、(c)所示,整理头发和双手放胸前的休息姿态都会被认为是手势.



(a) (b) (c)

Fig.1 The visual based detection and the failure cases

图 1 基于视觉的检测和检测失败的样例

1.2 基于音频的手势检测

本文利用语音端点检测进行基于音频的手势检测.用来计算短时能量的窗口为持续约 25ms 的音频.对计算出来的短时能量向量进行傅里叶变换和低通滤波,去除噪声,得到如图 2 中上方的曲线所示的音频短时能量图.定义每段的开始(图 2 中下方的竖直实线)和结束(图 2 中下方的竖直虚线),这些片段就被提取出来.但是,一些无意义的声音片段会干扰检测.

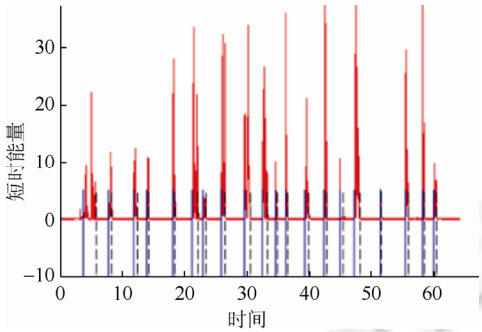


Fig.2 Segmentation based on audio. The curve on red is the short-time power after Fourier transformation

图 2 基于音频的分割.上方的曲线为经过傅里叶变换,去掉高频信息后的音频短时能量

1.3 融合音频和视觉的手势检测

可以发现,仅仅依靠视觉或者音频的手势检测在实际应用中并不鲁棒.但是,同样可以推断出基于视觉的手势检测的错误可以通过加入音频信息得到纠正,反之亦然.因此,融合基于视觉和音频的手势检测方法的动机和意义非常明显.为便于说明,基于独立音频和视觉检测的结果用两个 mask 表示,分别记为 m_a 和 m_v,其中手势帧标记为 1,非手势帧标记为 0.融合的过程是对这两个 mask 做合并和取舍操作,流程如图 3 所示.

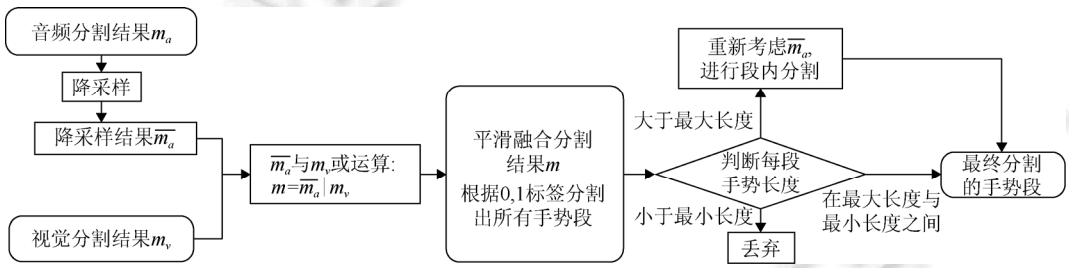


Fig.3 Flowchat of fused detection with audio and visual cues

图 3 音频和视觉检测融合流程图

2 手势分类

首先介绍音频和视觉特征的协方差矩阵的构造.在此基础上,提出一种在 Grassmann 流形下的距离度量方法.因为采样频率不同,需对视觉特征和音频特征独立地训练分类器,独立地分类并在结果层进行融合.用 α_{au} 和 α_{vi} 分别表示音频和视觉的分类结果融合权重.

2.1 构造协方差矩阵

首先提取手势的视觉特征.每一帧中左手和右手的区域由左右手骨架点所在的位置决定.提取该区域内手部图像的 HOG 特征,用 $\boldsymbol{f}_p$ 表示.骨架特征采用了 Wang 等人[9]提出的骨架对特征,用 $\boldsymbol{f}_s$ 表示.由于手势是上肢动作,因此只有头部、左右肘部和左右手部这 5 个骨架点参与骨架对特征的计算.图 4 显示了骨架对特征和手型特征组成的视觉特征.最后,用 $\boldsymbol{f}_v$ 表示每帧的连接手型和骨架的视觉特征,得到

$$f_v^i=\left[\boldsymbol{f}_p,\boldsymbol{f}_s\right]^i,i=1,...,K \tag{2}$$

其中,K 是一个动态手势视频的总帧数.

然后提取手势的音频特征.本文将梅尔倒频谱参数(12 维)、其一阶差分(12 维)和二阶差分(12 维)连接起来,构成了 36 维的音频特征,用 $\boldsymbol{f}_a$ 表示.音频和视觉特征各自构造协方差矩阵.为了便于本节的说明,将每个音(视)

频帧特征统一表示为 $\boldsymbol{f}$,其维数用 D 表示.一个手势序列的协方差矩阵可以由公式(3)进行计算,

$$C=\frac{1}{K-1}\sum_{i=1}^{K}(\boldsymbol{f}_i-\boldsymbol{\mu})(\boldsymbol{f}_i-\boldsymbol{\mu})^T \tag{3}$$

其中,K 代表手势片段的总帧数,μ 代表视频片段中所有特征的平均值.得到的协方差矩阵 C 是一个 $D\times D$ 的矩阵.

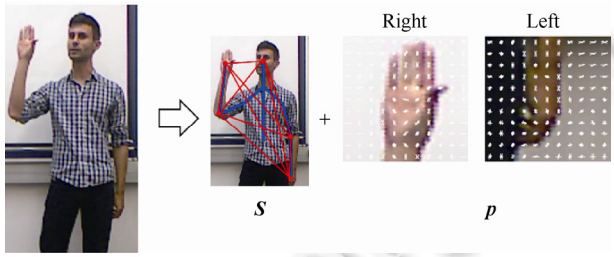


Fig.4 The visual feature. s: Pair-Wise skeleton feature. p: HOG feature

图 4 视觉特征提取.s 代表骨架对特征,p 代表手型的 HOG 特征

2.2 Grassmann流形上的度量

协方差矩阵是一个实对称半正定矩阵.只要在其对角线加上足够小的正实数,该矩阵就可以转化为对称正定(symmetric positive definite,简称 SPD)矩阵.通常,SPD 矩阵定义在黎曼流形上.为了保证计算速度,其距离通常近似地投影到欧式空间进行计算.为此,本文首次提出一种在 Grassmann 流形上更加精确地度量 SPD 矩阵距离的方法.首先,将 SPD 矩阵 C 进行 SVD 分解得到

$$C=Y\Sigma Y^T \tag{4}$$

其中 Y 是一个正交矩阵,它可以展开成一组列向量$[y_1,y_2,\ldots,y_D]$.选取前 $d(d\ll D)$个对应特征值最大的向量组成新的矩阵 $\overline{Y}=[\boldsymbol{y}_1,\boldsymbol{y}_2,...,\boldsymbol{y}_d]$.矩阵 $\overline{Y}$ 就可以看做一个理想的 R^D 空间的子空间.因此,可以用 Grassmann 流形上的度量来计算距离.矩阵 $\overline{Y}$ 在本文中被表示为(Grassmann covariance matrix,简称 GCM).距离度量的详细计算请见文献[11].分类器选用(kernel support vector machine,简称 Kernel SVM).其中两个 GCM 的 Kernel 被定义为

$$K(\overline{Y}_i,\overline{Y}_j)=\left\|\overline{Y}_i^T\overline{Y}_j\right\|^2 \tag{5}$$

3 实验结果与分析

3.1 数据库介绍

手势检测和识别算法的测试在 ChaLearn Multimodal Gesture Dataset [12]上进行.Chalearn 数据库是用 Kinect 采集的多人动态手势库,旨在实现非特定人的基于多模态特征的动态手势识别.采集的词汇集是 20 个常用的意大利手势.采集的数据包括语音、彩色图像、深度信息和人体遮罩.其中,测试者在做出相应手势的同时会用语言表达出该手势代表的意义.有 27 个测试者参与数据库的采集,一共采集到 956 个视频.每个视频持续时间约 1~2 分钟,包含 8~20 个数目不等的连续手势.其中,393 个视频(7 754 个手势)属于训练集,287 个视频(3 362 个手势)属于验证集,276 个视频(2 742 个手势)属于测试集.数据库的详细情况见文献[12].Chalearn 数据库的评价指标是根据编辑距离 LD(levenhstein distance)计算的.该评价系数用 LD 表示如下,

$$LD=(Ins+Del+Sub)/N \tag{6}$$

其中,Ins,Del,Sub 分别代表增加、删除、替换的误差个数,N 代表测试集中总的手势个数.为了更加详细地分析算法的表现,本文还引入了常用的召回率(recall)和准确率(precision)来辅助评价和分析.

3.2 测试结果

3.2.1 与其他方法的比较

Chalearn 数据库是在 2013 年的多模态手势识别挑战赛(Multi-Modal Gesture Recognition Challenge 2013)

上推出的.这次比赛的第 1、2、3、6 名的测试结果见表 1.第 6 名被作为对比是因为其提供了 Recall 和 Precision 的评测.本文对训练数据也进行了自动手势检测而没有直接使用竞赛方提供的人工分割.从表 1 中看到,本文的方法性能最优,其 LD 数值小于第 1 名 1 个百分点,并且 Recall 和 Precision 都高于第 6 名所提供的数值.具体地,在本文方法中,*Ins*,*Del*,*Sub* 的错误数量分别为 157,129,34.在最后的分类结果融合阶段,α_{au} 和 α_{vi} 分别为 0.3 和 0.7.在分类中,*d* 的取值为 10.算法能够在配备了 Intel i7 处理器和 32G 内存的机器上达到实时.

Table 1 Evaluations and comparisons on the ChaLearn dataset

表 1 ChaLearn 数据库上的测试评价和比较

方法	LD	Recall(%)	Precision(%)
IVA MM (Rank 1)	0.127 56	–	–
W WEIGHT (Rank 2)	0.153 87	–	–
ET (Rank 3)	0.168 13	–	–
LRS (Rank 6)	0.177 30	89.39	90.72
Ours	**0.117 30**	**94.02**	**93.07**

3.2.2 各模态特征独立测试

为了对比多模态特征对手势识别的性能,本实验用独立的特征进行手势建模和识别.其他实验设置不作改变.具体的识别结果见表 2.

Table 2 Evaluations with different features on the ChaLearn dataset

表 2 ChaLearn 数据库上各模态特征的测试评价和比较

特征	音频	骨架 S	手型 H	视觉(S+H)	所有特征
LD	0.452	0.509	0.366	0.279	**0.117 3**

可以发现,融合所有特征进行手势识别带来的 LD 分数的降低是非常明显的.在单独的特征中,手型特征的识别结果最好,骨架特征的识别结果最差.其原因是这 20 个意大利手语的骨架运动趋于一致,难以准确区分.相比之下,其手型变化则较大.

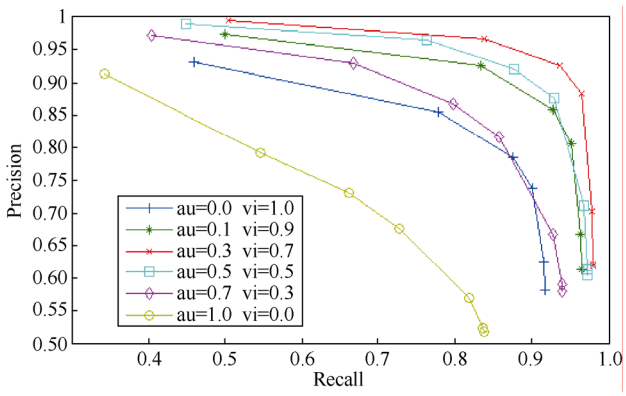


Fig.5 The Recall and Precision with different weights of visual and audio cues

图 5 在结果层融合中,不同的音频和视觉特征权重的 Recall 和 Precision

对于音频和视觉特征的识别结果在最后阶段的融合,本文通过改变它们的权重 α_{au} 和 α_{vi} 进行了全面的测试.不同权重下的识别结果用 Recall 和 Precision 评价,绘制的 PR (precision-recall)曲线如图 5 所示.可以发现,α_{au} 和 α_{vi} 分别为 0.3 和 0.7 时,Recall 和 Precision 的值皆最大.而 α_{au} 和 α_{vi} 分别为 1.0 和 0.0(即只依靠音频)时的结果最差,从另一个侧面说明提升基于音频特征的识别效果能进一步提高手势检测和识别的能力.

3.2.3 手势检测对结果的影响

本实验独立地测试基于音频和基于视觉的检测对识别结果的影响.在这组实验中,除了测试集的手势检测用独立特征以外,其模型训练阶段的检测所用的特征也做了相应的改变.在基于独立音频检测的实验中,训练数据所用的检测也仅仅依靠音频.同样地,在基于独立视觉手势检测的实验中,训练的数据所用的检测方法也仅仅依靠视觉.实验结果见表 3.基于音频的检测只达到了 0.231 0 的 LD 分数.基于视觉的检测也不能保证高的准确率.大部分检测误差是由于测试者的小动作和非默认的休息姿态引起.这导致了其 LD 分数比融合特征的手势检测高 7 个百分点.可见,融合音频和视觉的检测对提升手势识别结果的鲁棒性和识别率是显著的.

Table 3 Comparisons of different detection methds on the ChaLearn dataset

表 3 ChaLearn 数据库上的不同手势检测方法的比较

检测方法	LD	Recall (%)	Precision (%)
基于音频的检测	0.231 0	83.50	91.26
基于视觉的检测	0.187 0	86.32	80.81
融合手势检测	**0.117 3**	**94.02**	**93.07**

4 结束语

本文提出一种基于音频和视觉特征的手势检测和分类方法,能够在连续视频流中依次识别预定义的手势.在 Grassmann 流形下的距离度量保证了不同类的手势能够被高效地区分.实验结果表明,在 ChaLearn 数据库上,该算法的召回率和准确率都达到了 93%以上.融合音频和视觉特征带来的提升也在实验中得到了体现.后续工作将关注于提升音频特征对于检测和分类的贡献.

References:

[1] Pavlovic, VI, Sharma R, Huang TS. Visual interpretation of hand gestures for human-computer interaction: A review. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1997,19(7):677−695

[2] Erol A, Bebis G, Nicolescu M, Boyle RD, Twombly X. Vision-Based hand pose estimation: A review. Computer Vision and Image Understanding, 2007,108(1/2):52−73

[3] Elmezain M, Al-Hamadi A, Appenrodt J, Michaelis B. A hidden Markov model-based continuous gesture recognition system for hand motion trajectory. In: Proc. of the 19th Int'l Conf. on Pattern Recognition, Piscataway, NJ, 2008. 1−4.

[4] Yang HD, Sclaroff S, Lee SW. Sign language spotting with a threshold model based on conditional random fields. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2009,31(7):1264−1277.

[5] Wu J, Cheng J, Zhao C, Lu H. Fusing multi-modal features for gesture recognition. In: Proc. of the 15th ACM on ICMI. 2013. 453−460.

[6] Hadfield S, Bowden R. Generalised pose estimation using depth. Trends and Topics in Computer Vision, 2012. 312−325.

[7] Ahonen T, Abdenour H, Matti P. Face recognition with local binary patterns. In: Proc. of the European Conf. on Computer Vision, 2004. 469−481.

[8] Liwicki S, Everingham M. Automatic recognition of fingerspelled words in British sign language. In: Proc. of the Computer Vision and Pattern Recognition Workshops. 2009. 50−57.

[9] Wang J, Liu Z, Wu Y, Yuan J. Mining actionlet ensemble for action recognition with depth cameras. In: Proc. of the Computer Vision and Pattern Recognition. 2012. 1290−1297.

[10] Shotton J, Fitzgibbon A, Cook M, Sharp T, Finocchio M, Moore R, Kipman A, Blake A. Real-Time human pose recognition in parts from single depth images. Communications of the ACM, 2013,56(1):116−124.

[11] Hamm J, Lee DD. Grassmann discriminant analysis: A unifying view on subspace-based learning. In: Proc. of the 25th Int'l Conf. on Machine Learning. ACM, 2008. 376−383.

[12] Escalera S, Gonzalez J, Baro X, Reyes M, Lopes O, Guyon I, Athitsos V, Escalante H. Multi-Modal gesture recognition challenge 2013: Dataset and results. In: Proc. of the ICMI. 2013. 445−452.

王汉杰,男,博士,主要研究领域为计算机视觉,模式识别,手语,手势识别.

陈熙霖,男,博士,研究员,博士生导师,CCF 会士,主要研究领域为图像理解,计算机视觉,模式识别,图像处理.

柴秀娟,女,博士,助理研究员,CCF 会员,主要研究领域为计算机视觉,模式识别,人机交互.