采用SIFT-BoW和深度图像信息的中国手语识别研究

杨 全[1,2] 彭进业[2]

(西安文理学院软件学院 西安 710065)[1] (西北大学信息科学与技术学院 西安 710127)[2]

摘 要 将深度图像信息引入手语识别的研究,提出了一种基于DI_CamShift(Depth Image CamShift)和SIFT-BoW(Scale Invariant Feature Transform-Bag of Words)的中国手语识别方法。该方法将Kinect作为视频采集设备,在获取手语彩色视频的同时得到其深度信息;首先计算深度图像中手语手势的主轴方向角和质心位置,通过调整搜索窗口对手势进行准确跟踪;然后使用基于深度积分图像的Ostu算法分割手势并提取其SIFT特征,进而构建SIFT-BoW作为手语特征并用SVM进行识别。实验结果表明,该方法单个手语字母最好识别率为99.87%,平均识别率96.21%。

关键词 SIFT-BoW,DI_CamShift,深度图像,Kinect,手语识别

中图法分类号 TP311 **文献标识码** A

Chinese Sign Language Recognition Research Using SIFT-BoW and Depth Image Information

YANG Quan[1,2] PENG Jin-ye[2]

(School of Software Engineering, Xi'an University of Arts and Science, Xi'an 710065, China)[1]

(School of Information Science and Technology, Northwest University, Xi'an 710127, China)[2]

Abstract Introducing the depth image information into sign language recognition research, a Chinese sign language recognition method based on DI_CamShift (Depth Image CamShift) and SIFT-BoW (Scale Invariant Feature Transform-Bag of Words) was presented. It uses Kinect as the video capture device to obtain both of the color video and depth image information of sign language. First, it calculates spindle direction angle and mass center position of the depth image correctly tracks gesture by adjusting the search window. Second, an Ostu algorithm based on depth integral image is used to gesture segmentation, and the SIFT features are extracted. Finally, it builds SIFT-BoW as the feature of sign language and uses SVM for recognition. The experimental results show that the best recognition rate of single manual alphabet can reach 99.87%, while the average recognition rate is 96.21%.

Keywords SIFT-BoW, DI_CamShift, Depth image, Kinect, Sign language recognition

1 引言

聋人使用的手语是一种信息交流、传递思想的自然方法,在聋校教学和聋人交流方面起着非常重要的作用。手语是由手的形状、手腕和手臂的运动、与之对应的人脸表情、嘴唇读音形状以及其他身体姿态相结合的人体语言。中国手语包括手指语和手势语两类:手指语由30个手指字母构成基本单位,是用手指的变化和动作代表一个汉语拼音字母,并按照汉语拼音规则和顺序依次拼出词语构成的语言;手势语是通过指示性的形象手指姿态模拟事物形状和动作的自然手势。由于汉字本身包含的字词较多,因此使用手势很难全面、准确地将其表述完整。而手指语与拼音的拼写方式一致,可以表达很多专业术语以及抽象概念,具有简明易学、手势较少的特点。因此,手语字母的识别是手语识别中非常重要的一部分。根据教育部及中国文字改革委员会等单位公布的现行中国手语实施方案,中国手语中包括30个手语字母:26个单字母(A—Z)和4个双字母(ZH、CH、SH、NG),如图1所示。

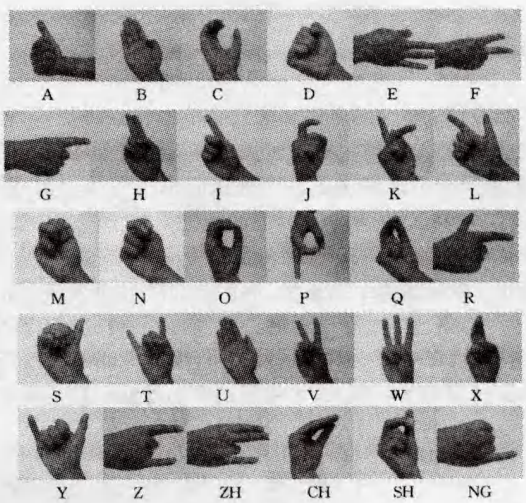


图1 手语字母表

到稿日期:2013-04-22 返修日期:2013-06-13 本文受国家自然科学基金项目(61075014),高等学校博士学科点专项科研基金(20116102110027)资助。

杨 全(1980—),女,博士生,讲师,主要研究方向为模式识别、数字图像处理,E-mail:yangquan1110@yeah.net;彭进业(1964—),男,教授,博士生导师,主要研究方向为数字图像处理。

手语识别研究包括两类：(1)基于计算机视觉的手语识别[1,2]；(2)基于佩戴式设备的手语识别。佩戴式位置跟踪器与数据手套由于价格较高，并且穿戴方式复杂，因此很难推广到实际应用中。随着视频采集设备的普及和应用，基于计算机视觉的手语识别以其自然、便捷的交互方式受到越来越多的关注，尤其是在硬件成本方面，该方式所需设备成本很低，因此非常适于普及应用，是目前手语识别的研究重点[3]。

哈尔滨工业大学的研究者利用圆形的轴对称和中心对称的几何特点，采用基于多尺度形状描述子的方法对静态中国手指语字母进行识别，识别率为94.6%[4]；在基于视觉的手语识别系统研究方面，中国科学院计算技术研究所的研究者采用CHMM对439个孤立手语词的平均识别率为92.5%[5]；哈尔滨工业大学的研究采用合成数据驱动算法对离线中国手语词进行识别，平均识别率为71.6%[6]。Deng等人[7]借助颜色手套简化手势分割，运用并行HMM识别192个美国手语词的识别率为93.3%。使用Harr特征并以AdaBoost作为识别分类器，Dinhet等人进行了24个英文手语的识别研究[8]。Silanon等人提出了根据手势运动分析识别泰文字母的方法[9]。Ong等人采用基于Sequential Pattern Trees的多级分类器方法对大词汇量单手语者手势序列和多手语者Kinect数据集进行识别[10]。文献[11]为了实现基于Kinect深度图像信息的手势轨迹识别，提出了一种基于HMM的手势轨迹识别方法并用于控制智能轮椅的运动。文献[12]采用Kinect摄像头采集手势深度图并转换为三维点云，根据深度信息过滤提取手势数据，在进行方向校正后统计手势数据中深度信息的区间分布特征并输入到SVM进行训练，实现了对数字手势1—5的识别，平均识别率达到95%。

本文采用Kinect作为手语视频采集设备，在获取彩色手语视频的同时得到相应的手语深度图像信息，在此基础上对传统的CamShift算法和Ostu算法进行了改进。通过提取手语手势的SLVW特征，构建其SIFT-BoW模型对手语字母进行识别，并通过实验验证了算法的准确性。

2 基于深度信息的手势跟踪与提取

在具有复杂背景的手语视频中提取手语手势作为下一步识别分析的目标，包括运动手势区域检测跟踪和手势的提取两个方面。

本文采用帧差的方法实现手语视频中手势区域的检测，通过改进的基于深度图像信息的DI_CamShift(Depth Image CamShift)算法确定手形区域，对手语手势进行跟踪，提出了一种基于深度图像信息的椭圆边界肤色建模方法；给出了基于深度积分图的二维Ostu算法，并采用改进的肤色分割算法提取运动目标区域的手势图像。

2.1 DI_CamShift跟踪

帧差法是背景减图法中的一种，其背景模型就是上一帧的图，所以使用帧差法进行运动手势检测不需要建模，速度较快。因为手语视频所处场景光照情况比较稳定，所以本文选用对缓慢变换光照不敏感的帧差法进行手语手势检测可以获得较好的效果，如图2所示。

图2 手势检测

由于传统CamShift算法是以颜色概率分布图为基础进行的目标跟踪，在复杂背景及目标位置发生突变时将导致跟踪失败。在进行手语手势跟踪的应用中，CamShift算法在反向投影过程中的全局运算增加了不必要的运算负担，因此会大大降低跟踪质量，主要原因是在亮度不强的背景区域，反向投影会产生噪点，在RGB转换到HSV空间的时候，亮度或饱和度不大的像素映射稳定性不高，会导致一些无关点也被当成目标区域。由于跟踪模型是建立在颜色模型的基础之上的，因此有可能让跟踪窗口中包括与目标无关的物体。在复杂背景下，使用CamShift算法跟踪手语手势时，视频中可能出现其他运动物体被误跟踪的情况，如果出现其他运动着的人手，则这种误判情况会非常明显。

Kinect采集的深度图像所包含的信息是与摄像机拍摄场景物体表面距离Kinect远近相关的数据，所以深度图像与彩色图像相比，能够更直观地体现物体表面的3D特征，不会受到颜色、亮度或阴影问题的干扰。根据深度图像的定义，可知其性质如下：(1)颜色无关性，与彩色图像不同，深度图像不会受光照、阴影或周围环境变化的影响；(2)深度图像灰度值变化的方向与视角中的Z方向一致，所以可以利用深度图像在可行的区域内重建三维空间。当出现遮挡或物体的重叠问题时，它对该问题的解决也可起到一定的作用。因为根据深度图像的特性，如果出现两个物体相互遮挡，则由其距离摄像头远近所产生的灰度值的分层情况，可对它们进行区分。只要设定一个阈值，将具有前后关系的物体分割开，就能够实现可见光图像无法解决的遮挡问题。

鉴于传统CamShift算法在彩色空间转换和运动手势跟踪方面的不足，本文使用图像的深度信息改进CamShift算法，得到Depth Image CamShift(DI_CamShift)算法。

对于深度图像$D(x,y)$，它的$(p+q)$阶二维原点矩M_{pq}的定义为：

$$M_{pq}=\sum_x\sum_y x^p y^q D(x,y),\ p,q=0,1,2,\cdots \tag{1}$$

$D(x,y)$为深度图像中(x,y)位置处像素的深度值。

定义$D(x,y)$的$(p+q)$阶中心矩μ_{pq}为：

$$\mu_{pq}=\sum_x\sum_y (x-\bar{x})^p (y-\bar{y})^q D(x,y) \tag{2}$$

则其二阶中心矩可以作为手语图像中手势的主轴。最大二阶矩和最小二阶矩的方向可用来确定手语手势的主轴，即长轴和短轴。手势主轴方向角θ可根据矩理论按式(3)求取：

$$\theta=\frac{1}{2}\tan^{-1}\frac{2\mu_{11}}{\mu_{20}-\mu_{02}} \tag{3}$$

式中，θ为主轴与坐标轴的夹角，其范围在$[-\frac{\pi}{4},\frac{\pi}{4}]$，如表1所列。

表 1　手势主轴与坐标轴的夹角

μ_{11}	$\mu_{20}-\mu_{02}$	θ
0	−	0
+	−	$-\pi/4<\theta<0$
+	0	0
+	+	$0<\theta<\pi/4$
0	0	0
−	+	$-\pi/4<\theta<0$
−	0	0
−	−	$0<\theta<\pi/4$

令手语手势 S 的主轴方向为 θ，则

$$S^2(\theta)=\frac{1}{n}[(S_1(\theta)-m)^2+(S_2(\theta)-m)^2+\cdots+(S_n(\theta)-m)^2] \tag{4}$$

其中，$m=\frac{1}{n}[S_1(\theta)+S_2(\theta)+\cdots+S_n(\theta)]$，$S_i(1,2,3,\cdots,n)$ 为从同一手语手势平均帧数中提取的目标主轴方向。

DI_CamShift 算法的具体步骤为：

(1)将整个深度图像设为搜索区域；

(2)采用帧差法检测视频中的运动人手区域，对 Search Window 进行初始化，定位其尺寸和位置；

(3)对 Search Window 区域内的部分，计算其深度直方图的概率分布；

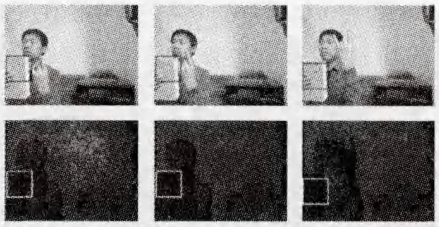
(4)分别计算深度图像中手势的长轴和短轴的方向 θ_1，θ_2；

(5)使用 MeanShift 算法计算窗口内深度手势图像的质心位置，根据质心位置和主轴方向 θ_1，θ_2 调整 Search Window 的尺寸；

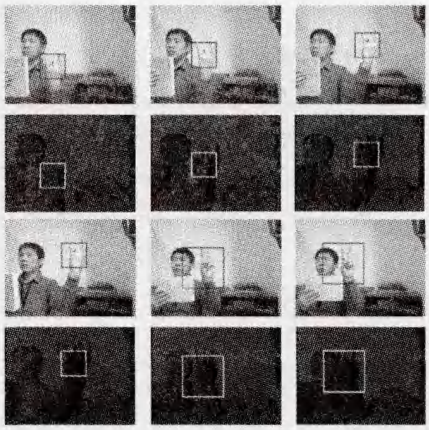
(6)对于下一帧手语视频图像，采用步骤(5)的 Search Window 的质心和尺寸，跳转至(3)继续运行；

(7)如果检测到多个运动目标，则目标手势为 HandGesture=Min$\{M_{00}(Obj_1),M_{00}(Obj_2),\cdots,M_{00}(Obj_n)\}$。因为距离摄像机远的物体具有较大的深度值，而手语识别中认为待识别手语手势是在打手语者身体之前，距离 Kinect 摄像机最近的目标物体，所以选取具有最小 0 阶矩值的 Search Window，该窗口内的手势像素具有最小的深度信息值的和，可确定为最前景目标手势区域。

如图 3(a)所示，视频中出现了一本黄色封面的书，由于 CamShift 只适于特定颜色目标的跟踪，而黄色较接近手语手势的肤色，所以在人手靠近黄色封皮的书时，CamShift 将跟踪窗口转移到了书的区域，将书误判为运动跟踪目标。并且，在人手运动远离书时，跟踪窗口始终停留在原处，未能对运动着的手部区域进行正确跟踪(彩色图像为手语视频中由 Kinect 彩色摄像头获取的彩色视频帧，下方对应的是 Kinect 深度摄像头同步拍摄获取的深度视频帧)。图 3(b)为相同场景情况下，DI_CamShift 对运动手势的正确跟踪。虽然受光线和拍摄角度影响，书皮的颜色在彩色视频帧中有时表现为深黄色，有时表现为浅黄色，但是深度信息不受上述因素影响，所以在人手靠近、远离书的运动过程中，DI_CamShift 始终将跟踪窗口定位于运动的人手区域，实现了准确的运动手势跟踪。



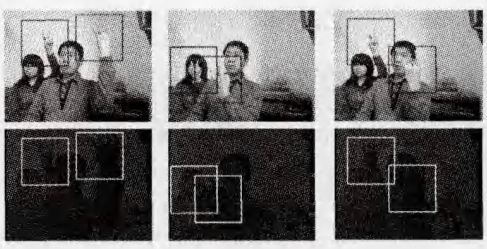
(a) CamShift 跟踪误判



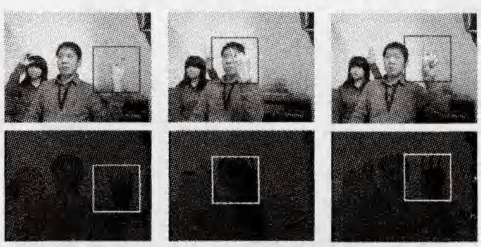
(b) DI_CamShift 抗干扰正确跟踪

图 3　类肤色物体干扰信息下 DI_CamShift 与 CamShift 手势跟踪比较

在进行手语手势跟踪时，认为手语手势位于人体的最前方，即当视频中出现多个运动着的人手的时候，借助 Z 方向的深度信息确定只有处于人的身体前端且距离摄像机最近的人手为跟踪目标。图 4(a)为视频中同时出现两个运动手势时 CamShift 的跟踪结果。虽然两个手势距离摄像头的前后位置不同，而且后方的人手为干扰因素，只有距离摄像头最近的打手语者的右手为待跟踪目标，但是由于它们都在运动，因此 CamShift 将两个运动手势同时定位且进行了跟踪。图 4(b)是 DI_CamShift 在手势深度信息基础上正确识别跟踪的目标手势。



(a) CamShift 无法区分视频中的多个手势



(b) DI_CamShift 始终追踪唯一的目标手势

图 4　DI_CamShift 与 CamShift 跟踪结果比较

一旦在深度手势视频中确定了跟踪窗口，就同步地将该

窗口绘制到彩色视频中对应的位置处，实现彩色视频中手势的跟踪。相同场景下，DI_CamShift 算法具有更好的跟踪效果，不会出现在跟踪中丢失手势目标的情况，也去除了距离摄像头较远的具有相同颜色信息的人手区域的误判跟踪。

2.2 手势分割提取

为了加快手势的分割提取计算速度，将深度图像的积分图像用于以类间方差作为准则函数的二维 Ostu 算法[13,14]，结合椭圆边界肤色模型，可得出一种改进的基于深度积分图像的 Ostu 算法。

(1)g 为椭圆边界手势肤色概率分布深度图，采用 3 * 3 的均值滤波算子对 g 进行滤波，可以得到其邻域平滑图像 f；通过遍历 g、f，寻找对应的像素并进行统计，其结果生成矩阵 d；最后计算 P(二维直方图矩阵)和它所对应的变换矩阵 P_i，P_j：

$$P(i,j)=\frac{d(i,j)}{M},M=\sum_{1}^{L}\sum_{1}^{L}d(i,j) \tag{5}$$

$$P_i(i,j)=(i-1)*P(i,j),P_j(i,j)=(j-1)*P(i,j) \tag{6}$$

(2)计算点(i,j)分别在 P、P_i 和 P_j 中的积分图：

$$\begin{aligned}G_0(i,j)&=\sum_{x=1}^{i}\sum_{y=1}^{j}P(x,y)\\ G_i(i,j)&=\sum_{x=1}^{i}\sum_{y=1}^{j}P_i(x,y)\\ G_j(i,j)&=\sum_{x=1}^{i}\sum_{y=1}^{j}P_j(x,y)\end{aligned} \tag{7}$$

(3)计算 S_B 和 S_T，它们分别为点(i,j)对应的类间离差矩阵和总离差矩阵：

$$S_B=\sum_{k=1,3}w_k*(u_k-u)(u_k-u)^{\mathrm{T}} \tag{8}$$

$$u=(u_i-u_j)^{\mathrm{T}}$$

$$S_T=\sum_{i=1}^{L}\sum_{j=1}^{L}p(i,j)*(x-u)(x-u)^{\mathrm{T}} \tag{9}$$

$$x=(i-1,j-1)^{\mathrm{T}}$$

其中，

$$u_i=G_i(L,L),u_j=G_j(L,L)$$

$$w_1=G_0(x,y),w_3=f_0$$

$$u_1=(G_i(i,j)/w_1,G_j(i,j)/w_1)^{\mathrm{T}}$$

$$u_3=(f_i/w_3,f_j/w_1)^{\mathrm{T}}$$

(4)计算 S_T-S_B，获得类内离差矩阵，计算 $\min\{|S_T-S_B|\}$((i,j)对应的准则函数值)，然后查寻所有数据点，获取最佳分割阈值。

图 5 为基于 Depth Image 和椭圆边界肤色模型的手语手势分割效果。

图 5 基于 Depth Image 椭圆边界肤色模型的分割

图 6 为在 DI_CamShift 跟踪窗口内进行的基于改进肤色分割算法的手势提取效果，作为深度距离最小的肤色物体，人手被较好地从背景中提取出来，未受视频中其他肤色物体或类肤色物体的影响。

图 6 复杂背景视频中手势的分割

3 SIFT-BoW 特征表示

BoW(Bag of Words)模型是信息检索领域常用的文档表示方法。在信息检索的时候，对于任何一个文档，BoW 模型都采取忽略它的单词前后顺序和语法、上下文等要素的方法，只是将其看作词汇的集合，文档中所有单词都被看作是独立而不依赖于其它任何单词的。也就是说，文档中任意一个位置出现的任何单词，都不受该文档语义的影响，是随机的。当用向量表示文档时，向量中每个元素表示词典中相关元素在文档中出现的次数，且未表达单词在原来句子中出现的次序。由此可知，假设存在一个巨大的文档集合 D，其中一共存储 M 个文档，提取出文档里面的所有单词，然后由这 N 个单词构成词典。假设用 BoW 模型描述文档，则每个文档都可以用一个 N 维向量来表示。

将 BoW 模型应用于手语识别，为了表征一幅手语图像，可以将该图像看作是一个文档，是若干个手语视觉词汇的集合，且不同视觉词之间不存在次序关系。由于手语图像中的手语视觉词汇不像文本文档那样直接显式存在，因此需要从图像中提取出相互独立的手语视觉词，对手语图像构造 BoW。该过程主要有 4 个步骤，如图 7 所示。

(1)特征检测：通过手语图像分割的方式获取手语手势的兴趣点；

(2)特征表示：采用图像的局部特征描述子来表示手语内容；

(3)生成视觉单词：把局部特征描述子表示的图像小区域量化生成手语视觉单词(Sign Language Visual Word，SLVW)；

(4)生成 BoW：统计一幅手语图像中的视觉单词频率，生成视觉单词频率直方图，用 BoW 模型来表示手语图像。

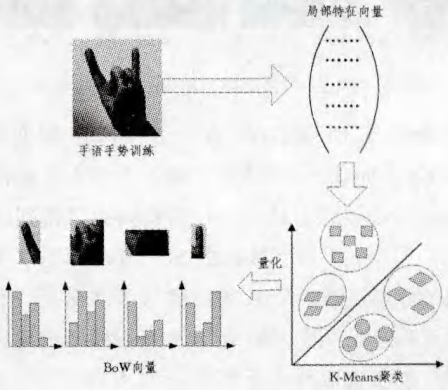


图 7 构造 BoW 模型的 4 个步骤

生成 BoW 后，通过使用 K-means 聚类算法，训练图像集上的所有特征，每一个聚类中心被定义为一个视觉单词，从而生成由 N 个视觉单词所组成的视觉单词词汇表。计算训练图像中图像块的特征与词汇表中的每个视觉单词所对应的欧式距离，与其中某个视觉单词距离最近的则被记录下来，不断重复以上过程，最终形成一组视觉单词频率统计直方图，生成视觉单词 BoW，用这组统计直方图代表该手语图像的特征[15,16]。SLVW 特征表示方式能够很好地解决图像局部特征点个数不同的问题，具有简单和只需要少量监督的优势[17]。

SIFT 的全称是 Scale Invariant Feature Transform，它是一种描绘局部特征的算子，通过在尺度空间搜寻极值点提取与旋转、尺度和位置无关的特征向量[18,19]。本文首先提取手语图像的 SIFT 特征作为图像中的手语视觉词汇，将所有的视觉词汇集合在一起，则每幅手语图像都由一个特征向量集合描述。然后，采用 K-Means 聚类算法构造手语视觉词汇表。SIFT 提取的视觉词汇向量之间根据距离的远近，可以利用 K-Means 算法将词义相近的词汇合并，得到 K 个聚类中心作为视觉单词，生成单词表中的基础词汇，再由视觉单词构成视觉词汇表，进而利用手语单词表中的词汇表示手语图像。利用 SIFT 算法，可以从每幅图像中提取多个特征点，这些特征点都可以用单词表中的单词近似代替，通过统计单词表中每个单词在图像中出现的次数，可以将图像表示成为一个 K 维数值向量。对每幅手语图像中的每个向量计算 N 个最近邻视觉单词，然后度量第 k 个 SLVW 的质量：

$$q(t_k)=\sum_{i=1}^{N}\sum_{j=1}^{M_i}\frac{1}{2^{i-1}}s(f_j,t_k) \tag{10}$$

式中，M_i 表示与视觉单词 t_k 第 i 接近的特征向量数，$s(f_j,t_k)$ 表示特征向量 f_j 与视觉单词 t_k 之间的相似度，本文采用 Squared Euclidean distance 进行度量。计算每个视觉单词的质量，如果满足

$$q(t_k)\leqslant h\max_{i=1,2,\cdots,K}(q(t_i)) \tag{11}$$

则保留，否则舍去。式(11)中，h 为视觉单词的质量阈值。图 8 分别为手语字母 M 和 W 的 SIFT 特征提取和 SIFT-BoW 特征构建。

(a) M的SIFT特征 (b) M的SIFT-BoW特征 (c) W的SIFT特征 (d) W的SIFT-BoW特征

图 8 手语字母的 SIFT 和 SIFT-BoW 特征

图 9 为基于 SIFT-BoW 的手语字母匹配实验，采用 Squared Euclidean distance 进行计算。大图为手语视频帧图像，左上角的小图为 DI_CamShift 窗口内跟踪得到的手势区域。图 9(a)为生成手语字母 A 的 SIFT-BoW 后，从手语库中随机选取的静态手语字母 A 的图像与手语视频帧中的字母 A 手势的正确匹配。图 9(b)—(f)为当视频中出现的手势是字母 A 时，依次选取静态手语字母 B、F、W、V、L 的 SIFT-BoW 与之匹配失败。图 9(g)(h)依次为手语字母 B、L 的正确匹配。通过实验结果比对分析可以发现，手语 SIFT-BoW 是一种非常有效的手语特征。

(a) 字母A正确匹配 (b) A与B无法匹配 (c) A与F无法匹配 (d) A与W无法匹配

(e) A与V无法匹配 (f) A与L无法匹配 (g) 字母B正确匹配 (h) 字母I正确匹配

图 9 基于 SIFT-BoW 的手语匹配

4 手语识别实验

实验运行环境：CPU：Intel Core i5-2300(双核、2.8GHz)；内存：4G；操作系统：win7 x64，安装. NET Framework 4.5，Kinect for Xbox 360 sensor 和 Xbox 360 Kinect AC Adapter/Power Supply；开发环境：vs2010 x64，Kinect SDK v1.7，OpenCV 2.4.4，OpenNI2.2 64-bit。

手语视频由 Kinect 采集，在白天自然光照、复杂背景下录制。由 2 位打手语者面向 Kinect 打出 30 个手语字母，视频包括打手语者的上半身区域和周围复杂环境场景。每个手语字母录制时长为 3 分钟，分别采集彩色视频和深度视频，两类视频的 30 个分段视频，共计 360 分钟。

本文实验首先使用 Kinect 采集手语视频图像，然后通过帧差法进行手语手势的检测和初步定位，并应用 DI_CamShift 算法进行手势跟踪。在手势分割提取时，用改进的肤色分割方法提取目标手势，并进行 SIFT-BoW 特征构建，最后采用径向基核函数(radical basis function，rbf)SVM 分类器进行训练[20]：

$$k(x,x_i)=\exp(-\gamma|x-x_i|^2) \tag{12}$$

实验从 360 分钟自然光照复杂背景手语视频中间隔采样，针对每位打手语者，每个手语字母手势分别提取 50 幅彩色图像和 50 幅深度图像，共计 6000 幅手语图像作为训练样本。从手语视频中非训练样本部分选取 30 个手语字母中的各 50 幅彩色图像和 50 幅深度图像作为测试对象，基于 rbf 的 SVM 分类器识别结果如表 2 所列，平均识别率为 96.21%。

表 2 手语字母识别结果

字母	识别率(%)	字母	识别率(%)
A	97.25	P	96.78
B	99.52	Q	93.58
C	98.88	R	96.13
D	98.54	S	97.52
E	98.84	T	98.14
F	98.69	U	95.21
G	95.24	V	99.87
H	94.26	W	98.67
I	96.48	X	93.22
J	88.53	Y	98.78
K	89.98	Z	95.56
L	97.47	ZH	97.12
M	92.14	CH	96.86
N	92.58	SH	95.68
O	98.99	NG	95.88

相同实验条件下，提取相同手语手势的不同常用表观特征，比对结果如表 3 所列。其中误判率为 SVM 对相似字母的误识别率。在相同手语库上的实验表明，SIFT-BoW 的方法在各项指标中均优于其他特征描述方法。

表 3　基于不同特征的手语识别方法比较

手语特征 \ 比较指标	平均识别率	最好识别率	误判率
Gray-Level Co-Occurrence Matrix	93.094%	98.97%	12.4%
Fourier Descriptor	89.74%	98.5%	11.37%
7Hu	80.2%	90%	22%
SIFT	94.52%	99.3%	9.8%
本文方法	96.21%	99.87%	5.25%

由于在视频手语手势跟踪过程中同时使用了彩色信息和深度信息，并且无需拍摄整个人体或转换为 2D 图像，因此在采用 DI_CamShift 手势追踪时避免了颜色、复杂背景和光照的影响，解决了 2D 手语识别时对环境背景要求较高，以及手指水平移动等造成的跟踪目标识别错误问题。在跟踪窗口内基于深度积分图像的 Ostu 算法实现了精确的人手分割提取，为 SIFT-BoW 模型的构建提供了准确的对象。SIFT 特征集虽然整体体现了手语手势的内容，但是仅为对该手势的一种不紧凑的表达，并不能体现手语手势的整体特性，并且 SVM 要求以相同维数的特征向量作为输入。SIFT-BoW 通过 K-means 聚类算法对手语手势的局部不变特征点进行聚类统计来表征手势，在保留 SIFT 特征较强的旋转、尺度、亮度不变性的同时，实现了手语字母的整体特征描述，获得了手势图像固定维数的表征向量，很好地解决了这一问题。

结束语　本文提出了一种基于 SLVW 的手语字母识别方法，其使用改进后的 DI_CamShift 算法通过计算深度图像中手语手势的主轴方向和质心位置调整 Search Window 的尺寸，实现了对手语手势的持续稳定跟踪；将深度积分图像应用于 Ostu 算法对手语手势进行分割提取；通过提取手语 SIFT 特征，将手语图像小区域量化生成其 SLVW 特征，然后使用 K-means 聚类算法得到视觉单词频率统计直方图，最终生成 SIFT-BoW。由于引入了手语的深度图像信息特征，使得识别过程不受颜色、光照和阴影的干扰。实验采用基于 rbf 的 SVM 分类器对 Kinect 采集的中国手语 30 个字母手势进行识别，获得了较高的识别率。

参 考 文 献

[1] Wachs J P, Kolsch M, Stern H, et al. Vision-Based Hand-Gesture Applications [J]. Communications of the ACM, February 2011, 54(2): 60-72

[2] Ren Zhou, Yuan Jun-song, Zhang Zheng-you. Robust Hand Gesture Recognition Based on Finger-Earth Mover's Distance with a Commodity Depth Camera[C]// The 19th ACM International Conference on Multimedia (MM'11). Scottsdale, Arizona, USA, November 28-December 1, 2011: 1093-1096

[3] Doliotis P, Stefan A, McMurrough C, et al. Comparing Gesture Recognition Accuracy Using Color and Depth Information[C]// Conference on Pervasive Technologies Related to Assistive Environments (PETRA). Crete, Greece, May 2011: 1-7

[4] 杨筱林，姚鸿勋. 基于多尺度形状描述子的手势识别[J]. 计算机工程与应用，2004(32)：76-78

[5] 张良国，高文，陈熙霖，等. 面向中等词汇量的中国手语视觉识别系统[J]. 计算机研究与发展，2006，43(3)：476-482

[6] 姜峰，高文，姚鸿勋，等. 非特定人手语识别问题中的合成数据驱动方法[J]. 计算机研究与发展，2007，44(5)：873-881

[7] Deng J W, Tsui H T. A two-step approach based on HMM for the recognition of ASL[C]// The 5th Asian Conference on Computer Vision. Melbourne, Australia, Jan 2002: 1-6

[8] Chen Qing, Georganas N D, Petriu E M. Real-time vision-based hand gesture recognition using Haar-like features[C]// Instrumentation and Measurement Technology Conference Proceedings. May 2007: 1-6

[9] Silanon K, Suvonvorn N. Hand motion analysis for Thai alphabet recognition using HMM [J]. International Journal of Information and Electronics Engineering, 2011, 1(1): 65-71

[10] Eng-Jon O, Nicolas C H P, et al. Sign Language Recognition using Sequential Pattern Trees[C]// The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Providence, Rhode Island, June 2012: 2200-2207

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

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

[13] 朱志亮，刘富国，陶向阳，等. 基于积分图和粒子群优化的肤色分割[J/OL]. http://www.cnki.net/kcms/detail/11.2127.TP.20130129.1543.016.html，2013-01

[14] 郎咸朋，朱枫，都颖明，等. 基于积分图像的快速二维 Otsu 算法[J]. 仪器仪表学报，2009，30(1)：39-43

[15] 王宇石，高文. 用基于视觉单词上下文的核函数对图像分类[J]. 中国图象图形学报，2010，15(4)：607-616

[16] 刘扬闻，霍宏，方涛. 词包模型中视觉单词歧义性分析[J]. 计算机工程，2011，34(19)：204-209

[17] 张秋余，王道东，张墨逸，等. 基于特征包支持向量机的手势识别[J]. 计算机应用，2012，32(12)：3392-3396

[18] Juan L, Gwun O. A Comparison of SIFT, PCA-SIFT and SURF [J]. International Journal of Image Processing, 2009, 3(4): 143-152

[19] Bastanlar Y, Temizel A, Yardimci Y. Improved SIFT matching for image pairs with scale difference. Electronics Letters [J]. 2010, 46(5): 107-108

[20] Lin C J. LibSVM: A library for Support Vector Machines [EB/OL]. http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2012-11