基于自适应希尔伯特扫描和词袋的图像检索

徐 墨,刘福岩,余梦婷
(上海大学 计算机工程与科学学院,上海 200444)

摘 要:提出一种自适应希尔伯特扫描方法用于解决图像检索中使用词袋模型丢失空间信息的问题。该方法通过分析特征在图像中的分布来计算在越来越精细的分辨率下每个希尔伯特路径的权重,从而为每张图像选择最优的扫描路径。探讨了基于希尔伯特扫描树的构建过程并对其优缺点进行了分析,该方法能够将图像特征的空间信息有效地加载到树的每个节点上。然后基于局部特征在图像空间的分布提出一种多层次的自适应希尔伯特扫描策略。得益于此方法,在之后为图像建立的基于希尔伯特树形结构上,物体的空间信息将被保存得更加准确,从而有利于快速重建物体轮廓。在 Caltech-256 数据库上进行了大量对比实验,实验结果表明该方法具有更高的检索准确率。

关键词:希尔伯特扫描;图像检索;词袋;特征表示
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2016)12-0017-05
doi:10.3969/j.issn.1673-629X.2016.12.004

Image Retrieval Based on Adaptive Hilbert Scan and Bag of Features

XU Mo,LIU Fu-yan,YU Meng-ting
(School of Computer Engineering and Science,Shanghai University,Shanghai 200444,China)

Abstract:One fundamental problem in large scale image retrieval with the bag-of-features is its lack of spatial information. An approach called adaptive Hilbert-scan depended on distribution of features in an image is proposed. This method computes weight of each Hilbert-scan at increasingly fine resolutions by analysis of feature distribution in the image,which is able to assign a suitable scanning path for each image. Hilbert-scan based tree structure is studied and its advantage ad disadvantage is analyzed. The method adds the spatial information of local features into each node of tree,furthermore a novel adaptive Hilbert-scan strategy with multi-level is designed,which is built on the distribution of features in image. Owing to merits of this method,spatial information of features will be preserved more precisely in Hilbert-scan based tree structures. Extensive experiments on Caltech-256 show the effectiveness of the method.

Key words:Hilbert-scan;image retrieval;bag-of-features;feature representation

1 概 述

近年来,大规模数据库中相似图像检索受到越来越多的关注。给出一张检索图像,目的是从数据库中检索出与其包含相同物体或场景的图像。一个大规模检索系统必须考虑三个因素:检索准确率、内存使用和效率。

目前大部分先进的检索技术都是基于词袋模型(Bag-Of-Features,BOF)[1]。BOF 模型的基本思想是将图像表示成一组视觉单词的直方图。基于 BOF 的检索系统首先需要提取每张图像的局部特征,例如最流行的 SIFT[2];然后对特征空间聚类生成字典;最终通过特征映射,每张图像可以用一个直方图向量来表示。然而 BOF 忽略了特征的空间信息,不能获取物体轮廓或者将物体从背景中分离出来。

为了解决这个问题,研究者们提出了一些 BOF 的衍生模型。例如文献[3]将图像分割成不同尺度下 $2^l \times 2^l$ 个小块,$l = 0,1,2$,再分别计算每个小块中 BOF 直方图向量,最后将 21 个小块中的向量连接起来形成最终的图像描述向量。尽管此方法简单可行,也取得了不错的效果,但它必须将图像分解成固定尺寸,导致其对严重混杂现象,几何变形(如旋转、尺度变化等)很敏感。文献[4]提出一种空间权重 BOF 方法,此方法通过降低背景特征权重,从而突出物体轮廓与位置的重要性。尽管该方法对于混杂背景具有鲁棒性,但它只适用于一些特定种类的图像。文献[5]将特征从二维空间以不同角度映射到一维空间,然后选取最具

收稿日期:2016-02-22 **修回日期**:2016-06-09 **网络出版时间**:2016-11-22
基金项目:国家自然科学基金面上项目(61471232)
作者简介:徐 墨(1989-),男,硕士,研究方向为图像处理与计算机图形学;刘福岩,副教授,研究方向为计算机图形学等。

代表性的表示方法作为最终的直方图描述向量。文献[6]对特征向量重新编码,将特征的空间信息加入投资向量中。此方法虽然有效,但字典维度太大,不适用于大规模图像检索。受文献[5]的启发,Hao 和 Kamata[7]将希尔伯特曲线与词袋模型相结合,为每张图像建立基于希尔伯特扫描的树形结构(Hilbert-Scan Based Tree,HSBT),特征空间信息通过一种聚类与过滤规则加载到树的每个节点,但该方法忽略不同扫描路径下生成的树形结构,可能导致空间信息误加载。

文中研究主要基于 HSBT[7] 结构,因为 HSBT 能有效地抓取物体轮廓并且压制背景信息,但其对于特征空间信息的加载可能存在错误。使用希尔伯特曲线对图像进行扫描时,必然导致图像中相邻两块区域在希尔伯特序列中分开。对一般图像而言,由于大量特征提取自图像中关键物体的局部外观,如果这些特征所处区域位于希尔伯特扫描的分裂处,必然造成一些空间上相关的特征在映射到一维空间后也被分开,一些融合错误必然产生在 HSBT 建立过程中,导致抓取的物体轮廓失真,或者不能有效地将物体从背景中分离出来,进而影响最终的检索准确率。于是,文中提出一种全新的自适应希尔伯特扫描策略用于为每张图像选择最适合它的希尔伯特扫描路径。该策略基于两个因素,第一是统计不同扫描路径下分裂出两个方块中包含的特征总数。由于一般图像中大部分特征都提取自关键物体,如果有一组相邻方块包含的特征点总数多于其他三组,这就意味着这两块区域很重要,不应该在一维空间中被分开,因此要避免使用该扫描路径。然而,如果这两块相邻区域中的大部分特征集中在其中一块中,那么经过希尔伯特扫描之后,二维空间中的相关特征在一维空间依旧离得很近,保证了特征的空间位置一致性。为了有效地融合这两个因素,加入一个权重参数用于控制它们的相对重要性。受希尔伯特扫描生成方法的启发,将此方法从特征的全局几何分布到局部几何分布进行实施,从而更加有效地运用特征的空间信息。

2 基于自适应希尔伯特扫描的词袋模型(AHS-BOF)

2.1 基于希尔伯特扫描的树形结构

假设一张图像的分辨率是 $m_1 \times m_2$,当提取了 SIFT 特征点之后,用希尔伯特扫描[8]将这些特征点映射到一维空间并将一维序列平均分割成很多子区域,因为希尔伯特扫描能够尽可能保存点的邻域[9]。由于图像尺寸不一定是正方形,于是提出一种伪希尔伯特扫描[10],它最重要的优点是像希尔伯特扫描一样能够尽可能保留每个点的邻域。因此,该算法有助于文中研究工作。首先用伪希尔伯特曲线扫描整张图像得到一个一维序列,用 $\{L,F,R\}$ 来表示。其中,L 表示特征点在图像中的坐标集;F 表示特征集;R 表示 L 分割后所有子区域的集合。在映射与分割之后,通过对这些子区域进行聚类操作从而建立 HSBT 结构。

每个子区域由两部分组成:标签与数据。HSBT 的第 i 层中第 j 个区域用 R_j^i 表示。数据包含四部分:区域中特征的个数(n_j^i)、区域的重心(g_j^i)、特征点集合(F_j^i)、区域的聚类中心(c_j^i)。重心通过计算特征的坐标得来,聚类中心用模糊 C-means 的方法对特征集合计算可得。

接着,主要探讨区域间是如何进行聚类生成 HSBT 结构的。树的第 i 层的所有区域将其表示为 $R^i = \{R_j^i \mid i = 1,2,\cdots,v(i)\}$。那么第 i 层的数据集可以表示成 N^i, G^i, F^i, C^i,$N^i = \{n_j^i \mid j = 1,2,\cdots,v(j)\}$,$G^i = \{g_j^i \mid j = 1,2,\cdots,v(i)\}$,$F^i = \{f_j^i \mid j = 1,2,\cdots,v(i)\}$,$C^i = \{c_j^i \mid j = 1,2,\cdots,v(i)\}$。区域间聚类主要有三个步骤:初始化、重要区域选择、区域融合。

初始化:首先将线性序列 S 平均分割成 $v(1)$ 个子区域,分解因子为 δ,那么,$m_1 \times m_2$ 大小的图像就被平均分割成 $v(1) = m_1 \times m_2/\delta$ 个不规则的小块。由此可以统计每个子区域中的特征个数并过滤没有特征的空白区域,进而可以得到 G', F', C'。

重要区域选择:在区域融合之前,首先选择重要的子区域,具体细节如下:

(1)需要根据子区域中的特征个数对其升序排序。排序之后,$R^i = \{R_j^i \mid j = 1,2,\cdots,v(1)\}$ 可以表示成 $R^i = \{R_{\alpha(j)}^i \mid \alpha(j) \in [1,2,\cdots,v(i)]\}$。其中,$n_{\alpha(1)}^i > n_{\alpha(2)}^i > \cdots > n_{\alpha(v(j))}^i$。

(2) $\sum_{j=1}^{\alpha(s)} n_{\alpha(j)}^i > \mathrm{Th} \times M$; $\sum_{j=1}^{\alpha(s-1)} n_{\alpha(j)}^i < \mathrm{Th} \times M$。

其中,M 为图像中特征点的个数;Th 为阈值,0<Th<1。

因此 $R_{\mathrm{main}}^i = \{R_{\alpha(1)}^i, R_{\alpha(2)}^i, \cdots, R_{\alpha(s)}^i\}$,$1 \leqslant \alpha(s) \leqslant v(i)$。

(3) $v(i+1) = \alpha(s)$; $R_{\mathrm{rest}}^i = R^i - R_{\mathrm{main}}^i$。

区域融合:当选出重要区域与非重要区域之后,需要将非重要区域融入重要区域。例如,有三个相邻的子区域:R_x^i 和 R_y^i 是重要区域,R_z^i 是非重要区域。那么应该与哪块区域融合。Hao 和 Kamata[7] 提出一种新的融合准则(见式(1)),该融合准则主要源于万有引力定律。

$$\frac{n_x^i}{|g_x^i - g_y^i|} > \frac{n_z^i}{|g_z^i - g_y^i|} \tag{1}$$

通过以上三个步骤可以为一幅图像建立 HSBT 结

构,如图 1 所示。

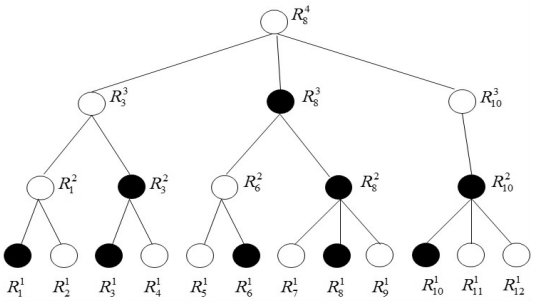


图 1　图像的 HSBT 结构

HSBT 旨在将二维空间中的特征映射到一维空间,为每张图像生成一个树形结构。无需任何标记或手动操作,通过一定的聚类规则,特征的空间信息都被加载到树的每个节点中。

在区域融合过程中,单一并且固定的扫描路径可能会产生大量的融合误差。目标是保证从关键物体中提取出来的特征在被映射到一维空间之后尽可能离得很近。在区域融合过程中,来自关键物体的特征会更快被融合到一起,从而有利于快速且准确地恢复关键物体轮廓,避免被不必要的噪声特征干扰。下一节依据特征在图像空间中的分布提出一种全新的自适应希尔伯特扫描策略。该策略可以为每张图像选择合适的希尔伯特扫描路径,从而在建立 HSBT 结构时减少融合误差并快速重建关键物体的轮廓。

2.2　自适应希尔伯特扫描

对于一般图像,假设对其提取 SIFT[2] 特征点,那么大部分特征都是从其关键物体中提取出来的。在建立 HSBT 时,包含绝大多数特征的区域被选为重要区域,通过多次区域选择与融合从而建立树形结构。目标是确保这些重要区域在希尔伯特序列中尽可能离得很近,从而有效地减少融合误差,快速重建重要物体或者场景的轮廓。对于一张图像存在四种扫描方式(如图 2 所示,其中黑色点表示特征点),也会产生四种不同的线性序列。

图 2　四种希尔伯特扫描

在图 2 中不难发现,第二与第三块区域包含图像中大部分特征,因此这两块区域在建立 HSBT 时会被选为重要区域。如果使用图中第四种扫描路径,那么块二与块三在一维序列中将被分开很远。因而,很多不相关的特征点(比如从人和道路上提取出来的特征点)会夹杂在块二与块三之间,导致很多不相关特征融入块二与块三中,产生许多融合误差,不利于重建重要区域轮廓。为了避免上述不好的扫描路径,基于特征的分布情况提出一种全新的自适应希尔伯特路径选择策略。

上面提到过,如果某区域包含了图像中大部分特征,那么该区域就是重要区域,如图 3 中的块二与块三。因此,影响路径选择的第一个因素是在分裂边缘(图 3 中的黄色线条表示分裂边缘)两边的子块中所包含特征的个数。所以,第一个因素的公式表示为:

$$w_{1S} = n_A + n_B / n_C \tag{2}$$

其中,n_A 和 n_B 分别表示分裂边缘两边子块各自包含特征的个数;n_C 表示整个图像包含特征的个数;s 表示第 s 个扫描路径。

如图 2 中所示,第三个扫描路径中分裂边缘两边子块包含更多特征,那么它们就是重要区域,不应该被分开。但这时会产生一个问题:如果图 3 中 A 和 B 两块中的特征集中分布在其中一块中,即使这两块在一维序列中被分开,绝大多数特征仍然在一维空间离得很近,因而不会产生很多融合误差。为了解决这个问题,文中又提出了另一个影响路径选择的因素:

$$w_{2S} = \min(n_A, n_B) / \max(n_A, n_B) \tag{3}$$

如果接近 1,那么 A 和 B 中所包含的特征个数几乎相同。当特征被映射到一维空间时,大部分特征将被分离得很远,从而产生大量融合误差;相反,接近 0 时,表明大部分特征集中在其中一个方块中,特征在一维空间中仍然距离很近,降低了融合误差产生的可能性。此时很难判断哪个因素对于最后的路径选择更重要,为了将这两个因素有效结合,引入一个权重系数 λ 来控制两个因素的相对重要性。公式如下:

$$W_S = (1 - \lambda) \cdot w_{1S} + \lambda \cdot w_{2S} \tag{4}$$

由文献[7]可知,对于一个矩形空间,一般需要经历数次分裂才会形成最终的希尔伯特扫描。例如一个大小为 4×4 的矩形需要两次分裂才会生成最后的希尔伯特扫描(为基本形,不能再进行分裂,如图 4)。一般的图像尺寸从几十乘以几十到几千乘以几千,很显然仅仅计算一次分裂远远不足以评判哪个扫描路径更好。比如,当和很相近时,宏观上几乎无法判断哪个路径更好(如图 2 中的第二种和第四种)。因此,需要将目光投向特征在每个小块中的分布,即更细粒度的分布。需要从宏观和微观两个视角分析特征分布状况。

在此基础上提出了一种多层次方法,见式(5)。

$$W_s = \sum_{i=1}^{k} \sum_{j=1}^{l} [(1-\lambda) \cdot w_{1s}^{ij} + \lambda \cdot w_{2s}^{ij}] \cdot \frac{c^{ij}}{c} \quad (5)$$

其中,i 表示第 i 次分裂;j 表示在第 i 次分裂中第 j 个子块;k 表示分裂次数;l 表示每次分裂下子块的个数;c 表示一张图像中所有特征的个数;c^{ij} 表示第 i 次分裂中第 j 个子块中特征点的个数;c^{ij}/c 表示这个小块相对于整张图像的权重。

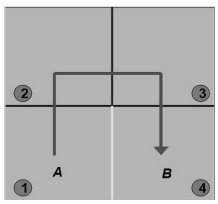


图 3 第一层希尔伯特扫描

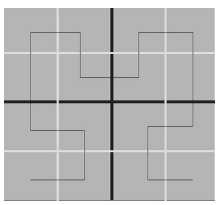


图 4 4 * 4 希尔伯特扫描

如果一个子块包含的特征点多,那么它相对比较重要,它所分配到的权重值也应该越大。

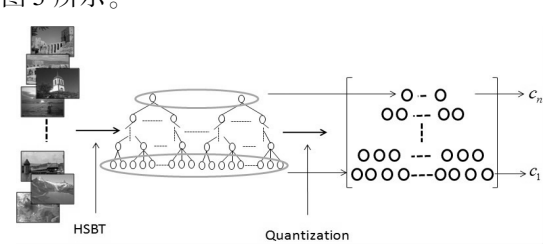
上述方法不仅考虑了特征的全局几何分布,也考虑了其局部几何分布情况。由粗到细,从宏观到微观,全方位地考虑了特征的空间信息。

2.3 基于自适应希尔伯特扫描的词袋模型

在为每一张图像选择了合适的路径并建立 HSBT 之后,将 BOF 模型与 HSBT 相结合,这一部分主要有两个步骤:字典生成和特征编码。

(1)字典生成。

假设一个训练集中有 Q 张图像,那么最终的特征集合可以表示成 $\varphi = \{X_1^1, X_1^2, \cdots, X_1^{h_1}; X_2^1, X_2^2, \cdots, X_2^{h_2}; \cdots; X_Q^1, X_Q^2, \cdots, X_Q^{h_Q}\}$。其中,$h_Q$ 表示图像 Q 生成的 HSBT 高度。实验中采用文献[11]提出的多层字典生成法来设计字典。在该方法中,树的每一层节点 $X^i = \{X_1^i, X_1^i, \cdots, X_Q^i\}$ 都会生成一个字典 V^i,$i = 1,2,\cdots,H$。其中,$H = \max(h_1, h_2, \cdots, h_Q)$。因此,可以获得一个多层次字典:$\{V^1, V^2, \cdots, V^H\}$。多层次字典生成过程如图 5 所示。



万方数据多层字典的生成示意图

(2)特征编码。

在字典生成之后,所有的局部特征必须映射到字典中的每个视觉单词上。目前有许多字典映射的方法,例如向量量化(Vector Quantization,VQ)、稀疏编码[11](Sparse Coding,SP)、局部约束线性编码[12](Locality-constrained Liner Coding,LLC)和局部软分配编码[13](Local Soft-assignment Coding,LSC)等。VQ 是一对一映射方法,即每一个局部特征映射到离它最近的视觉单词上。但这没有考虑到视觉单词的歧义性,通常会带来量化误差。为了解决这一问题,提出了 SP、LLC、LSC 等方法,这些方法都将局部特征映射到多个视觉单词上,并在图像分类上取得了非常好的效果。然而这些方法或多或少都存在一些缺陷,其中最严重的缺点就是时间复杂度过大,而大规模图像检索问题对于检索速度要求很高。为了在检索准确率与检索速度寻求一个平衡,最终还是选择了 VQ 作为最后的特征编码方法。

对于一张图像,其树形结构的高度为 h,如果 h 大于 H,只对其前 H 层进行量化;如果 h 小于 H,只使用其前 h 层。最终,一张图像可以被表示成一个多层次直方图。编码规则如下:

$$g_t(\varphi_j^i(k)) = T_j^i = [t_1^i, \cdots, t_u^i, \cdots, t_{S_i}^i] \quad (6)$$

$$t_u^i = \begin{cases} 1, u = \arg\min_b \| \varphi_j^i(k) - v_b^i \| \\ 0, \text{otherwise} \end{cases} \quad (7)$$

其中,$\varphi_j^i(k)$ 为第 i 层上第 j 个节点中的第 k 个局部特征;v_b^i 是 V^i 中的一个视觉单词,$b = 1,2,\cdots,S_i$。

每一层的直方图生成之后,使用最常用的 TF-IDF 方法对其进行加权。最终,将每一层的直方图向量连接起来生成图像特征描述向量。在计算图像相似度时,采用文献[7]中的 CHI-Square 方法。

3 实 验

在 Caltech-256 上检验此方法。Caltech-256[14] 由加州理工学院于 2007 年建立,其中包含 256 个类别,每个类别至少包含 80 张图像,图像总数为 30 607。每个类别的图像都是由人工手动挑选出来的。图像类型涉及不同场景、光照、角度等条件下的自然与人工物体。实验采用 mAP(mean Average Precision)作为评价准则。为了训练字典,随机从每个类别中选取 50 张图像作为训练集,训练方法采用 K-means。然后再从每个类别中选 5 张图像作为测试集。设 δ 为 500,Th 为 0.8。和文献[7]一样,提取树的前 10 层,字典大小设置为 10 k,20 k,50 k,100 k。

图 6 显示了不同 λ 下几种方法的对比。很明显,第二层 AHS-BOF 的 mAP 优于其他几种方法。这也证

明了文中的设想:多层次能够有效抓取特征的空间信息。图中的 AHS-BCF-level2 几乎都在 HS-BOF 之上,这也表明该方法可能不会为图像选择最适合它的扫描路径,但一定不会产生不利影响。在第二层,该方法相比于 HS-BOF 提高了大概 3% 的准确率。

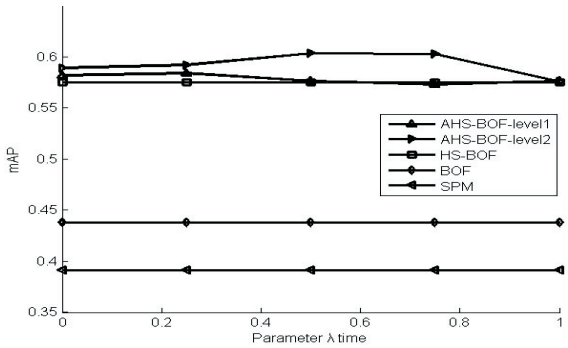


图 6　不同 λ 下的 mAP

表 1 显示了在不同字典大小下 AHS-BOF 与其他几种方法的对比。为了还原前人的结果,所有的实验参数与文献[7]一样。当字典大小为 100 k 时,其大小几乎等于特征的个数,因而直方图向量几乎不具有辨识力,因此,选择任何扫描路径对最终的检索结果都不会产生显著影响。所以,AHS-BOF 与 HS-BOF 的检索结果几乎相同。

表 1　不同字典大小下 mAP 的比较

Word size	BOF	SPM	HS-BOF	AHS-BOF
10 k	0.438	0.391	0.576	0.604
20 k	0.541	0.437	0.625	0.653
50 k	0.573	0.472	0.635	0.657
100 k	0.604	0.499	0.595	0.596

表 2 比较了不同分裂层数下 mAP 的大小。很明显第二层的结果最好,第三层结果有所下降。这表明,过度重视特征的细节信息会将一个统一的物体割裂开,从而丢失物体全局外观。

表 2　不同分裂次数下 mAP 的比较

L(分裂次数)	AHS-BOF($\lambda=0.5$)	HS-BOF
1	0.576	
2	0.604	0.575
3	0.592	

4　结束语

文中提出的自适应希尔伯特扫描策略能够为每张图像选择合适的扫描路径,减少了 HSBT 中的融合误差,降低了物体重建所需树的层数,节省了物体重建时间,有效地解决了 BOF 模型丢失特征空间信息的缺点。然而该方法对于复杂语义或者场景(卧室、客厅等)的图像处理能力不够,并且不同数据库最优的权重参数 λ 也不同。未来工作中,会尝试更抽象的特征提取方法,并设计一种动态参数调节方法。

参考文献:

[1] Sivic J, Zisserman A. Video Google: a text retrieval approach to object matching in videos[C]//Ninth IEEE international conference on computer vision. [s. l.]: IEEE, 2003: 1470-1477.

[2] Lowe D G. Object recognition from local scale-invariant features[C]//Proceedings of the seventh IEEE international conference on computer vision. [s. l.]: IEEE, 1999: 1150-1157.

[3] Lazebnik S, Schmid C, Ponce J. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories [C]//IEEE computer society conference on computer vision and pattern recognition. [s. l.]: IEEE, 2006: 2169-2178.

[4] Marszaek M, Schmid C. Spatial weighting for bag-of-features [C]//IEEE computer society conference on computer vision and pattern recognition. [s. l.]: IEEE, 2006: 2118-2125.

[5] Cao Y, Wang C, Li Z, et al. Spatial-bag-of-features[C]// IEEE conference on computer vision and pattern recognition. [s. l.]: IEEE, 2010: 3352-3359.

[6] McCann S, Lowe D G. Spatially local coding for object recognition[M]//ACCV 2012. Berlin: Springer, 2012: 204-217.

[7] Hao Pengyi, Kamata S. Hilbert scan based bag-of-features for image retrieval[J]. IEICE Transactions on Information and Systems, 2011, 94(6): 1260-1268.

[8] Hilbert D. Über die stetige Abbildung einer Linie auf ein Flächenstück[J]. Mathematische Annalen, 1891, 38: 459-460.

[9] Jagadish H, Faloutsos C, Saltz H. Analysis of the clustering properties of the Hilbert space-filling curve[J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 13(1): 124-141.

[10] Zhang J, Kamata S, Ueshige Y. A pseudo-Hilbert scan for arbitrarily-sized arrays[J]. IEICE Transactions on Fundamentals, 2007, 90(3): 682-690.

[11] Yang J, Yu K, Gong Y, et al. Linear spatial pyramid matching using sparse coding for image classification[C]//IEEE conference on computer vision and pattern recognition. [s. l.]: IEEE, 2009: 1794-1801.

[12] Wang J, Yang J, Yu K, et al. Locality-constrained linear coding for image classification[C]//IEEE conference on computer vision and pattern recognition. [s. l.]: IEEE, 2010: 3360-3367.

[13] Liu L, Wang L, Liu X. In defense of soft-assignment coding [C]//IEEE international conference on computer vision. [s. l.]: IEEE, 2011: 2486-2493.

[14] Griffin G, Holub A, Perona P. Caltech-256 object category dataset[R]. California: California Institute of Technology, 2007.