基于词频统计特征和 GVP 的大规模图像检索算法研究

刘 宇 邓 亮 郭耕辰 冯良炳

（中国科学院深圳先进技术研究院 深圳 518055）

摘 要 针对传统的 GVP（Geometry-Preserving Visual Phrases）图像检索算法计算量大、时间复杂度高且不适合处理大规模图像检索等缺点，文章提出了 FSF-GVP（Frequency Statistics Feature-Geometry-Preserving Visual Phrases）算法，该方法将词频统计特征和 GVP 算法相结合，使用 GVP 排序算法对词频特征统计后的相似结果集进行排序，忽略不相似结果集，极大地提高了检索效率。实验结果表明，FSF-GVP 在保证检索准确性的前提下，提高了检索效率，适用于实时大规模图像检索。

关键词 图像检索；词袋模型

中图分类号 TP 751 **文献标志码** A

Image Retrieval Using Feature Word Frequency Statistics of Geometry-Preserving Visual Phrases

LIU Yu DENG Liang GUO Gengchen FENG Liangbing

(*Shenzhen Institutes of Advanced Technology*, *Chinese Academy of Sciences*, *Shenzhen* 518055, *China*)

Abstract Traditional GVP (geometry-preserving visual phrases) image retrieval algorithm is not suitable for handling the large-scale image retrieval because of its high time complexity. In this paper, FSF-GVP (frequency statistics feature-geometry-preserving visual phrases) algorithm, which combined word frequency statistic characteristics and GVP algorithm, was proposed. FSF-GVP algorithm counts visual word frequency characteristics of an image to be searched and image database to get similar result set and dissimilar result set. Then FSF-GVP algorithm uses the GVP algorithm to sort the similar result set, which improves the retrieval efficiency. The experiment results on Oxford 5K dataset show that FSF-GVP is suitable for the large-scale real-time image retrieval on the premise of ensuring the accuracy of retrieving result and improving the retrieval efficiency.

Keywords geometry-preserving visual phrases; bag of words

收稿日期：2013-12-29

基金项目：国家自然科学基金项目（61070147），深圳市科技研发资金基础研究计划（JC201105190951A）。

作者简介：刘宇，硕士研究生，研究方向为图像检索和软件架构；邓亮，硕士研究生，研究方向为模式识别和图像处理；郭耕辰，硕士研究生，研究方向为软件架构；冯良炳（通讯作者），博士，研究方向为智能计算与智能控制、多视角三维重建、网络服务组合和增强现实，E-mail：lb.feng@siat.ac.cn。

1 引　言

图像检索在工业领域应用潜力巨大，近几年关于图像检索的研究引起了广泛关注。在图像检索领域，词袋模型(Bag of Words，BOW)方法被广泛使用。BOW 算法借鉴了 Web 文本搜索的 Bag of Word 思想，将图像分成一些视觉单词用于表达图像语义。BOW 首先由 Video Google 提出[2]，由 Li 等[3]改进 BOW 模型，应用于单一物体及多类物体的识别和分类，并创建了统一图像测试集 Caltect 101 和 Caltech 256。词袋模型[4]基于 SIFT(Similar Local Features，SIFT)点的描述基础上计算得到视觉词汇表，能够较好地描述目标的统计信息，也在文本分类领域取得了很好的结果。但是该特征是一个乱序的统计特征，并不能反映目标内部之间特征所在位置的相互关系。实验证明，词袋模型在目标识别方面有较好的识别结果。为了提高检索精确度、效率和排序结果，在词袋模型思想基础上，有很多的算法被陆续提出。Vocabulary Tree 数据模型[5]和最相似紧邻算法被有效地用于建立大型词典，软匹配(Soft Matching)[6]和汉明切入(Hamming Embedding)[7]被用来解决词袋模型的硬量化问题。Query expansion[8]作为后期步骤重排序以期获得高的召回率。除此还有很多改进的方法，如减少倒排表使用内存等。但是图像检索算法因为其计算量大、时间复杂度高而不适合处理大规模图像检索。其中一个最主要的限制是忽略了视觉单词的空间几何信息。

因为视觉单词本身有多重歧义性，所以空间几何信息的引入对于视觉表达非常重要。空间几何信息通常在后处理中使用，如随机抽样一致(Random Sample Consensus)[9]或者近邻特征一致(Neighboring Feature Consistency)[2]。因为几何验证方法通常计算量非常大，所以他们只被安排在了初始排序的前几个图像。金字塔匹配算法(Spatial Pyramid Matching)[10]虽然在空间上有所改进[11]，但是缺乏变换的灵活性。空间加权(Spatial-Bag-of-Features)算法通过改变空间直方图的顺序来解决金字塔匹配算法的可变性问题。对于每个视觉单词来说，空间直方图按照出现次数最多的方式被重排序，这种思想推进了词袋模型和金字塔匹配算法的发展。

GVP(Geometry-Preserving Visual Phrases)[1]是基于计算视觉单词同时出现次数进行得分排序的算法。对于特定空间的几何信息，一个 GVP 是一组视觉单词。所以，不同的 words 或者不同的空间布局对应不同的 GVP。GVP 被用来编码空间几何信息，也可以编码更多其他种类的信息。Zhang 等[1]发现将图像分为 6×7 的大块，当同一块中两个图像之间出现相同的 SIFT 特征时，取相同特征数的组合。当组合数取 2 时对于一般结果来说最好。Wang 等[12]基于 GVP 算法，提出提高组合数的取值时检索效果更好。但是 GVP 的性能一直不够好，且因为计算量大而不适合处理大规模图像检索。为了准确、高效地处理大规模图像检索，本文提出 FSF-GVP(Frequency Statistics Feature-Geometry-Preserving Visual Phrases)算法。算法首先统计数据库中图像与被搜索图像的词频特性，得到相似结果集和不相似结果集，再对相似结果集使用 GVP 算法进行排序，从而得到最终搜索结果。

本文首先对 FSF-GVP 进行具体描述，然后再将该算法与其他算法在时间和效率上进行对比分析。

2 FSF-GVP 图像检索算法

2.1 算法描述

2004 年，Lowe[13]总结已有的基于不变量技术的特征检测方法，正式提出了一种基于尺度空间，对图像平移、旋转和缩放，甚至仿射变换保持不变性的图像局部特征，即 SIFT 特征。

词袋模型基于 SIFT 特征，根据被搜索图像，在数据库图像中寻找同一类图像并返回图像序列。这种算法首先在文本处理领域获取了巨大的成功，它使用概率语义分析模型挖掘文档集中的主题信息，采用非监督学习方法，从底层的文本特征中获得语义。BOW 以其快速、高效性在文本分类问题中获得了巨大的成功[14-17]。关于分析文本与图像分类之间的关系，可以将两个问题进行类比分析，理解为一副图像中包含了很多视觉“单词”。从文本文档中可以得出词语—文档共生矩阵，类似的，从图像分析中可以得出图像的词语—文档矩阵。所以可以按照不同的视觉主题对图像进行分类。图 1 说明了文本分类与图像分类关系。

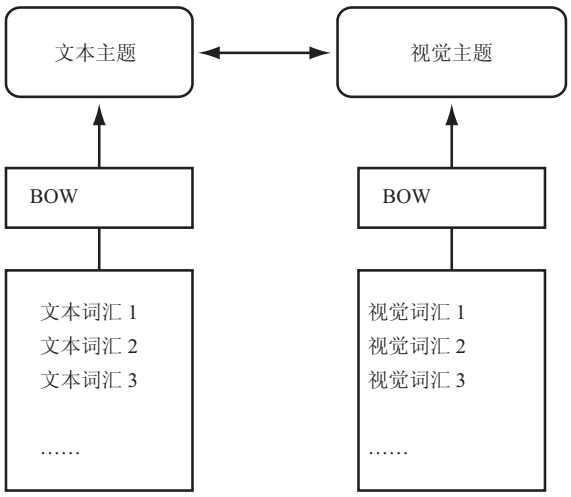


图 1 文本分类与图像分类的关系

Fig. 1. Relationship between text classification and image classification

GVP 算法补充了 BOW 算法，对 BOW 算法返回的图像序列进行排序，得到搜索数据库图像后的序列。GVP 算法将图像分为 $M=A\times B$ 的大块，计算每块中所有视觉单词同时出现的概率。当数据库图像某块中相同的某几个视觉单词同时出现时，就给数据库图像相似度总分加相应的权值分数。

GVP 算法描述如下：

(1) 初始化数据库中所有图像每个块(每个图像共 M 块)的分数为 0。每个块表达了视觉单词的空间几何信息。

(2) 对于被搜索图像 m 中的每个视觉单词 j，通过倒排表检索包含视觉单词 j 的所有图像 id 和视觉单词 j 出现的次数。对在图像 i 中出现的每个视觉单词 j，将它的空间几何位置与被搜索图中 j 的空间几何位置相减得到每块的分数。

$$S_{i,\,xd-xj,\,yd-yj}=S_{i,\,xd-xj,\,yd-yj}+1 \tag{1}$$

其中 (x_d, y_d) 是图像 i 中的 j 所处的空间几何位置，(x_j, y_j) 是被搜索图像 m 中 j 所处的空间几何位置。S_i 是图像 i 的分数。

(3) 遍历每个图像 i 的每块 m，计算 length 为 k 的 GVP 所得分数。对上一步算得所有图像的每块总分，使用组合方法求得每块的新分数。并将每块新分数加和得到一张图像的总分数。

$$S_i'=\sum_{m\geqslant k}\binom{S_{i,m}}{k} \tag{2}$$

(4) 对所有图像分数排序得到所需搜索图像序列。

GVP 算法还有一些重要的部分，如在上述算法中加入词频重要性权值[1]，加入对最后总分的规范化[1]，对 length 取值的选择[12]，对词频权值的快速选择[18]等。

依据相似图像的特征总是与几何位置相关的特性[12]，FSF-GVP 首先统计数据库中图像与被搜索图像视觉单词空间的词频特性得到相似结果集和不相似结果集，再对相似结果集使用 GVP 算法，利用视觉单词空间几何布局，对图像进行排序。

假设，有 n 张待搜索图像，对所有的图像的所有 SIFT 特征使用 K-means 方法划分类别，得到 c 个类。

FSF-GVP 算法步骤如下：

(1)对要搜索的图像，首先求出 SIFT 特征，并根据之前分类的方法将这些 SIFT 分类。取种类中出现次数多的前 k 个类别作为索引。设索引为 j，正在对比的图为 i，i 中 j 出现的次数为 $occur$，则每个图的分数为：

$$\sum_j S_{i,j}=\sum_j S_{i,j}+occur \tag{3}$$

(2)按照桶排序方式，将分数高于设定阈值的图像放入相似结果集列表，其他图像放入不相似结果集列表。

(3)对相似结果集图像列表的前 m 个图像进行 GVP 计算分数，再进行排序并得到最终搜索结果；

过程解释如下：首先计算被搜索图像出现的次数，取出视觉单词数量最多的 k 个类别作索引，然后对所有图像的视觉单词相同索引的出现次数进行简单相加，以相加后的和作为总分，使用桶式排序，把高于设定阈值的图像放入相似结果集列表中，低于设定阈值的图像放入不相似结果集列表中。这样对候选集进行了预筛选，极大缩小了搜索空间。对相似结果集图像列表取 m 个图像并使用 GVP 进行排序，得到最终搜索结果。

2.2　算法性能分析

GVP 算法首先计算被搜索图像的 SIFT 特征，然后对每个图像的所有 SIFT 类进行扫描，按照索引号进行加分，时间复杂度为 $O(c\cdot n)$。将所有图像分成 6×7 的格子，计算格子内符合 length 的组合数记入总分，最后对所有图像总分进行排序得到搜索结果，时间复杂度为$O(c\cdot n)+O(\log n)$。总的时间复杂度为：

$$O(c\cdot n)+O(c\cdot n)+O(\log n) \tag{4}$$

FSF-GVP 算法首先计算被搜索图像的所有 SIFT 特征，并取前 k 个分类作为索引，然后对每个图像与索引相同的 SIFT 出现次数进行相加得到初始分数，采用类桶式排序将分数高于阈值的放入列表中，时间复杂度为 $O(k\cdot c\cdot n)$。对相似结果集图像列表，对他们进行 GVP 排序，时间复杂度为 $O(c\cdot m\cdot n)+O(c\cdot m\cdot n)+O(\log(m\cdot n))$。总的时间复杂度为

$$O(k\cdot c\cdot n)+O(c\cdot m\cdot n)+O(c\cdot m\cdot n)+O(\log(m\cdot n)) \tag{5}$$

3　实　验

3.1　数据集和评估标准

Oxford 5K 数据集[9]已经成为一种评估标准。它包含了 11 种不同的 Oxford 里程碑和其他的扰乱选项，由 Flickr 提供 5062 图像分辨率。

本文使用主集合的平均准确率(Mean Average Precision，MAP)[9]来评估所有实验的性能。在相同的实验环境下，用执行时间来评估算法速度。并计算每个图像搜索的精确度求平均值来生成平均准确率的分数。

3.2　实验参数与结果对比

本文使用 128 维的 SIFT 描述子进行描述，使用 K-means 算法对 SIFT 描述子进行聚类，共 50000 个类。一般实验中，k 取 30%～40%，m 取 20%～30%。本文 k 取 30%，m 取 20%。

图 2 描述了 Oxford 5K 数据库上，GVP 算法在不同字典大小、不同深度 length 下的取值结果。

图 2 表明，词典越大，深度 length 的最佳值越高。因为深度 length 越高，要求在小范围内与原被搜索图相同的特征越多。而当 length 过大，筛选条件太过严格，搜索平均准确率下降。所有图像无论相似与否得分指标都很低，无法区分图像谁更相似，所以最后曲线都向下，MAP 降低。

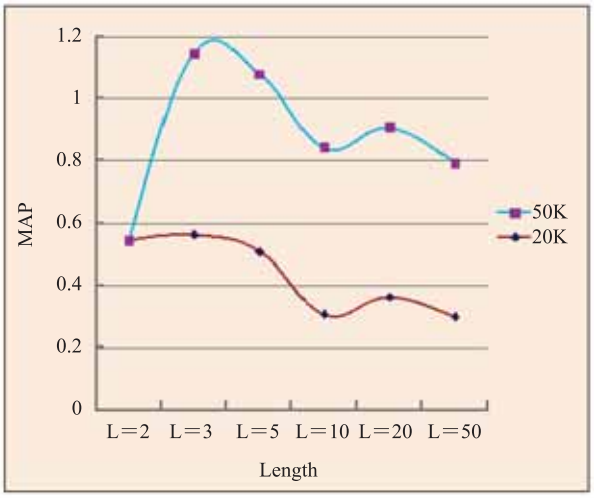


图 2 GVP 在不同词典、不同 length 下的性能比较

Fig. 2. GVP's performance under different dictionary sizes and different lengths

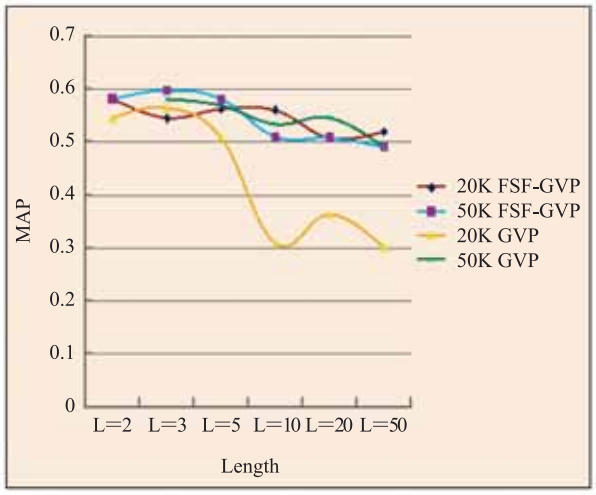


图 3 FSF-GVP 与 GVP 在词典为 20K 与 50K 和 length 不同下的效果比较

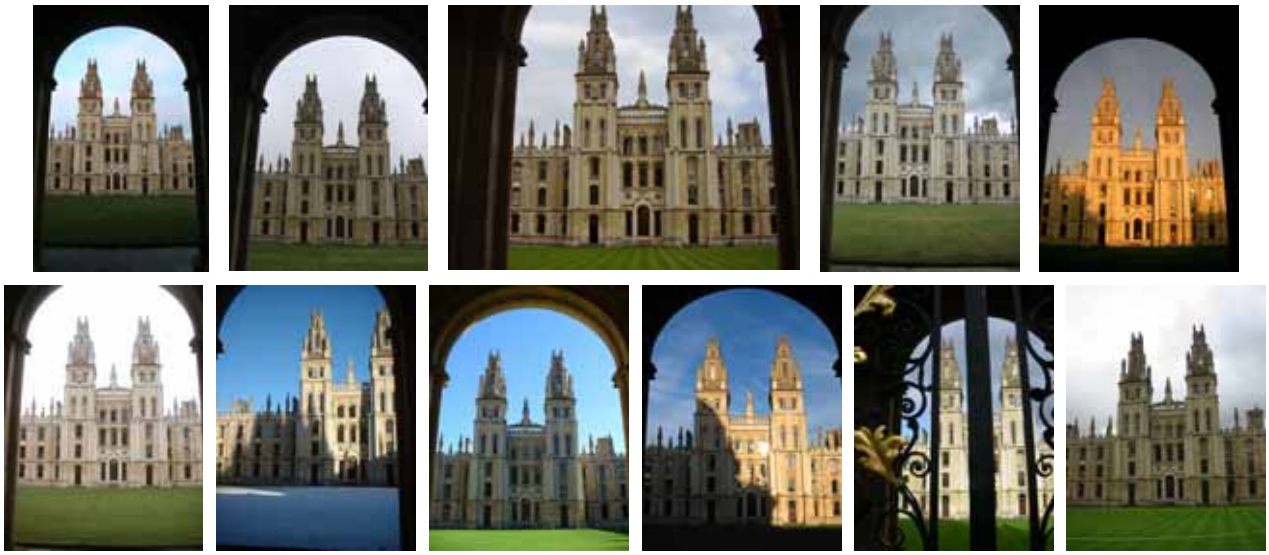
Fig. 3. Comparison between FSF-GVP and GVP of different dictionary sizes and lengths

图 3 比较了 GVP 与 FSF-GVP 排序后前 50 个图像的平均精确度。

图 3 的实验结果与图 2 比较相似。性能上几乎无变化。当 length 相对较大时，图中 FSF-GVP 算法的变化比 GVP 相对迟钝，即对 length 变化的耐受性更好。这是因为 FSF-GVP 的排序所用总分包含了两部分，一部分是数据库图像与被搜索图出现相同 SIFT 类别的次数作为分数，另一部分是 GVP 计算排序所用的分数。Length 只影响 GVP 计算排序所用的分数而不影响前一部分的分数，所以 FSF-GVP 对 length 的变化相对迟钝，即当 length 相对较大时平均搜索准确率下降速度较 GVP 更小。

文中 GVP 和 FSF-GVP 图像排序前 10 名平均精度结果表明，GVP 算法与 FSF-GVP 算法在图像排序前 10 名的平均精度上几乎无差别，其中，GVP 平均准确率为 0.870，而 FSF-GVP 为 0.868。因为对于排名非常靠前的图像，他们与被搜索图非常相似。对于 FSF-GVP 算法而言，越相似的图像，分数第一部分即相同 SIFT 类别的出现次数越高，并且第二部分 GVP 的分数也会越高。所以 FSF-GVP 与 GVP 的前 10 张图像搜索平均准确率基本相同。

图 4 为 GVP 和 FSF-GVP 的搜索结果。使用图 4 两幅图像各自的第一个图像作为被搜索图像，数据库图像中包含原图。实验中，GVP 与 FSF-GVP 都能在将原图排在第一位，并且前 10 的搜索精度无差别。



（a）GVP 搜索结果

（b）FSF-GVP 搜索结果

图 4　GVP 和 FSF-GVP 的搜索结果

Fig. 4.　The search results of GVP and FSF-GVP

4　结　论

为准确、高效的处理大规模图像检索，本文在 GVP 算法的基础上，提出了 FSF-GVP 算法。首先统计数据库中图像与被搜索图像的词频特性得到相似结果集和不相似结果集，极大缩小了搜索空间，再对相似结果集使用 GVP 算法进行排序。Oxford 5K 数据库上的实验表明排序后前 10 张图像的检索精度与 GVP 算法无差别。FSF-GVP 对于 length 能接受的波动范围比 GVP 要大很多。FSF-GVP 较 GVP 更快速，更适合于大规模的图像检索。

参 考 文 献

[1] Zhang YM, Jia Z, Chen T. Image retrieval with geometry-preserving visual phrases [C] // IEEE Conference on Computer Vision and Pattern Recognition, 2011: 809-816.

[2] Sivic J, Zisserman A. Video google: a text retrieval approach to object matching in videos [C] // International Conference on Computer Vision, 2003: 1470-1477.

[3] Li FF, Fergus R, Perona P. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories [C] // Conference on Computer Vision and Pattern Recognition Workshop, 2004: 178.

[4] He X, Zemel RS. Learning and Incorporating Top-down Cues in Image Segmentation [M]. Springer Berlin Heidelberg, 2006: 338-351.

[5] Nister D, Stewenius H. Scalable recognition with a vocabulary tree [C] // IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006: 2161-2168.

[6] Philbin J, Chum O, Isard M , et al. Lost in quantization: improving particular object retrieval in large scale image databases [C] // Conference on Computer Vision and Pattern Recognition, 2008: 1-8.

[7] Jegou H, Douze M, Schmid C. Hamming embedding and weak geometric consistency for large scale image search [C] // Proceedings of the 10th European Conference on Computer Vision: Part I, 2008: 304-317.

[8] Chum O, Philbin J, Isard M, et al. Total recall: automatic query expansion with a generative feature model for object retrieval [C] // Proceedings of the IEEE International Conference on Computer Vision, 2007: 1-8.

[9] Philbin J, Chum O, Isard M, et al. Object retrieval with large vocabularies and fast spatial matching [C] // IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2007: 1-8.

[10] 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, 2006: 2169-2178.

[11] Weiss Y, Torralba A, Reigus R. Spectral hashing [C] // Advances in Neural Information Processing Systems, 2008: 2169-2178.

[12] Wang FY, Zhang SW, Li HP, et al. Image retrieval using multiple orders of Geometry-preserving Visual Phrases [C] // International Conference on Image Analysis and Signal Processing, 2012: 1-5.

[13] Lowe DG. Distinctive image features from scale-invariant keypoints [J]. International Journal of Computer Vision, 2004. 60(2): 91-110.

[14] Grosky WI, Zhao R. Negotiating the semantic gap: from feature maps to semantic landscapes [C] // Theory and Practice of Informatics Lecture Notes in Computer Science, 2001: 33-52.

[15] Hare JS, Sinclair PAS, Lewis PH, et al. Bridging the semantic gap in multimedia information retrieval: top-down and bottom-up approaches [C] // Mastering the Gap: from Information Extraction to Semantic Representation, European Semantic Web Conference, 2006: 6894624.

[16] Zhang YJ. Visual Informsfen Retrieval Based on Content [M]. Science Press, 2003.

[17] Zhao R, Grosky WI. Narrowing the semantic gap-improved text-based web document retrieval using visual features [J]. IEEE Transactions on Multimedia, 2002, 4(2): 189-200.

[18] Wang FY, Zhang SW. Image retrieval using accurate approximatedinverse document frequency of geometry-preserving visual phrases [C] // International Conference on Audio, Language and Image Processing, 2012: 914-918.