

文章编号:1001-9081(2013)03-0667-03

doi:10.3724/SP.J.1087.2013.00667

基于 Bag-of-words 和 Hash 编码的近似重复图像检测算法

王誉天^{1*}, 袁江涛², 秦海权¹, 刘 鑫¹

(1. 公安部第一研究所, 北京 100048; 2. 天津市公安局北辰分局, 天津 300400)

(*通信作者电子邮箱 ytwang99@126.com)

摘要:针对近似重复图像检测的传统算法存在检测效率和准确率不够高的缺点,提出了基于 Bag-of-words 和哈希编码的近似重复图像检测算法。该算法首先利用 Bag-of-words 把一幅图像表示成一个 500 维的特征向量;然后,利用主成分分析(PCA)和尺度不变特征转换(SIFT)进行特征降维,并利用 Hash 编码技术对特征进行编码;最后,利用动态距离度量技术实现近似重复图像的检测。实验结果表明,利用该算法进行近似重复图像检测是完全可行的,在准确度和查全率之间做到了较好的平衡,查准率可达 90% ~ 95%,查全率可达 70% ~ 80%。

关键词:近似重复图像; Bag-of-words; 主成分分析; 哈希编码; 动态距离度量

中图分类号: TP301.4 **文献标志码:**A

Algorithm of near-duplicate image detection based on Bag-of-words and Hash coding

WANG Yutian^{1*}, YUAN Jiangtao², QIN Haiquan¹, LIU Xin¹

(1. The First Research Institute, Ministry of Public Security, Beijing 100048, China;

2. Beichen Branch, Tianjin Municipal Public Security Bureau, Tianjin 300400, China)

Abstract: To solve the low efficiency and precision of the traditional methods, a near-duplicate image detection algorithm based on Bag-of-words and Hash coding was proposed in this paper. Firstly, a 500-dimensional feature vector was used to represent an image by Bag-of-words; secondly, feature dimension was reduced by Principal Component Analysis (PCA) and Scale-Invariant Feature Transform (SIFT) and features were encoded by Hash coding; finally, dynamic distance metric was used to detect near-duplicate images. The experimental results show that the algorithm based on Bag-of-words and Hash coding is feasible in detecting near-duplicate images. This algorithm can achieve a good balance between precision and recall rate: the precision rate can reach 90% – 95%, and entire recall rate can reach 70% – 80%.

Key words: near-duplicate image; Bag-of-words; Principal Component Analysis (PCA); Hash coding; dynamic distance metric

0 引言

通常近似重复图像^[1]是由源图像通过某些变换得到的,一般可以产生近似重复图像的变换包括平移、缩放、旋转、图像色调的变化、添加文字、格式变化、分辨率变化等,除此之外还包括从不同视角拍摄的同一场景的图像,图 1 为经过了几种变化的近似重复图像。随着图像硬件和软件处理技术的快速发展和广泛应用,使得图像数据的采集、创作和存储成本日趋低廉化,每天都有数以万计的图像数据产生和发布,这些图像数据又通过不同的工具进行编辑、转换等操作变成其他的多个版本,然后通过网络进行分发,近似重复图像的检测已经变得越来越迫切。借助于近似重复图像检测,可以通过更改的图像为用户检测出原始图像,不但可以为用户提供高质量的新闻图片,而且可以进行版权保护,打击非法转载行为。

文献[2–7]采用全局特征进行近似图像检测,全局特征对视角变化,部分遮挡,光照变化和几何变换等表现得不够鲁棒,全局特征的检测速度较快,但是对近似重复图像检测效果比较差。目前越来越多的研究人员倾向于采用基于尺度不变特征变换(Scale-Invariant Feature Transform, SIFT)和主成分分析(Principal Component Analysis, PCA)的局部特征,文献[8]以及采用视觉单词直方图进行近似图像检测,但这两种

方法在检测近似区域较小的图像时效果不好。文献[9]采用基于贪婪树的外部支持向量机进行近似重复图像的聚类,该方法应用贪婪树将外部支持向量机聚类推广到多类聚类,同时使用模型将同现的视觉单词映射到潜在语义空间中的同一方向上,该方法取得了较好的效果,但是贪婪树最优化分解的过程需要花费很长时间。文献[10]提出了一种基于近似重复图像匹配的图书检索方法,特征提取采用的是改进 SIFT,虽然取得了较好的检测结果;但是检索库的规模小,而且只限定在图书,推广能力值得怀疑。基于局部特征的检测算法虽然可以比全局特征取得较好的检测效果,但是难以在检测准确性和检测效率之间做到一个平衡。

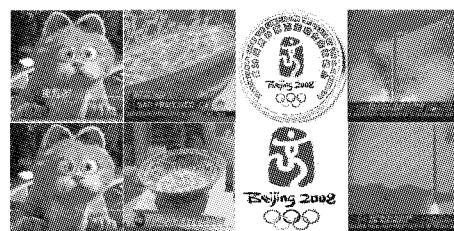


图 1 近似重复图像

为了兼顾检测的准确度和检测效率,本文提出了基于 Bag-of-words^[11–12]和 Hash 编码的近似重复图像检测技术。

收稿日期:2012-09-10;修回日期:2012-11-02。

作者简介:王誉天(1975–),男,陕西咸阳人,工程师,硕士,主要研究方向:电子与计算机测试;袁江涛(1976–),男,甘肃天水人,工程师,硕士,主要研究方向:计算机网络;秦海权(1981–),男,湖南永州人,工程师,硕士,主要研究方向:计算机信息安全、数据鉴定;刘鑫(1980–),女,山东济宁人,工程师,硕士,主要研究方向:计算机信息安全。

该方法使用 Bag-of-words 对一幅图像进行语义级的表示,采用 Hash 编码和动态距离技术实现近似重复图像的快速和精确检索,具体检测流程如图 2 所示。



图 2 基于 Bag-of-words 和哈希编码的近似重复图像检测流程

1 基于 Bag-of-words 的图像特征表示

基于 Bag-of-words 模型最初被用在文本分类中将文档表示成数字矢量。它的基本思想是假定对于一个文本,忽略其词序和语法、句法,仅仅将其看作是一些互相独立的词汇的集合。简单说就是将每篇文档都看成一个袋子,然后看这个袋子里装的是些什么词汇,将其分类。类似地 Bag-of-words 也可用在图像特征表示方面。基于 Bag-of-words 的图像表示模型方法一般包括 3 部分:1)利用 SIFT 提取图像的局部特征;2)通过聚类量化图像的局部特征,构建视觉单词;3)构建图像的视觉词频次直方图。采用 Bag-of-words 的实现过程如图 3 所示。

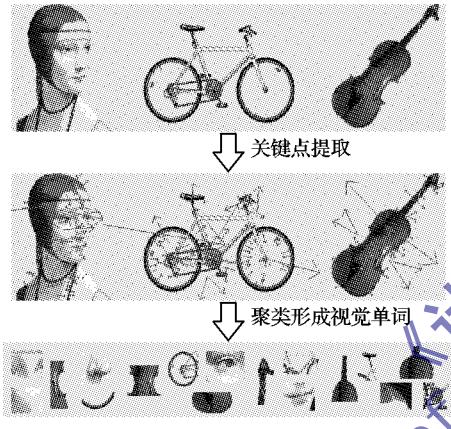


图 3 Bag-of-words 的实现过程

1.1 SIFT 特征提取

SIFT 尺度不变特征转换提取的是图像的局部特征,该特征对旋转、尺度缩放和亮度变化保持不变性,对视角变化、放射变换和噪声也保持一定程度的稳定性。SIFT 特征的提取包括 5 个过程:1)构建尺度空间;2)检测空间极值点;3)精确确定极值点;4)分配关键点方向;5)描述特征。一幅图像经过 SIFT 后,会产生很多的局部特征点,每一个点都是 128 维的特征向量。

1.2 构建视觉单词

SIFT 视觉特征到视觉单词的映射可以通过 k-means 聚类来实现,最终的聚类中心就是要得到的视觉词,聚类中心的数量就是码本矢量的大小。k-means 是一种简单的聚类算法,因其理论可靠、收敛速度快而被广泛应用。

k-means 算法采用迭代更新的思想,先将要参与聚类的图像库的 SIFT 特征数据载入内存;然后,随机选择 K 个对象对聚类中心初始化 c_1, c_2, \dots, c_k —— 初始化过程;再对剩下的每个对象 $x_i (i = 1, 2, 3, \dots, n)$ 根据其与各个簇中心的距离将它赋给最近的簇(m 是特征数据的维数)—— 分配过程:

$$\|x_i - c_j\| = \min_{1 \leq j \leq k} \sqrt{\sum_{l=1}^m (x_{il} - c_{jl})^2}; 1 \leq l \leq m \quad (1)$$

然后重新计算每个簇的均值作为下一次迭代的聚类中心—— 更新类中心:

$$c_j = \frac{1}{N_j} \sum_{x_i \in s_j} x_i \quad (2)$$

其中 N_j 为第 j 个簇 s_j 中的对象数目。分配和更新类中心两个过程要不断重复,直到所有类中心都不再变化为止,最后得到的类中心就是 Bag-of-words 模型所需要的视觉词,本文聚类中心取为 500 个。

1.3 构建图像的视觉词频次直方图

该过程要统计每幅图像中每个 SIFT 矢量与所有视觉词的距离,若其距离某个视觉词最近,就将该视觉词对应的 bin 高度加 1,直至将所有 SIFT 描述子矢量都分配完毕为止,这样每一幅图像都能用一个视觉词序列大小的直方图表示,如图 4 所示。这样每一幅图像都可用一个 500 维的特征矢量来表示。

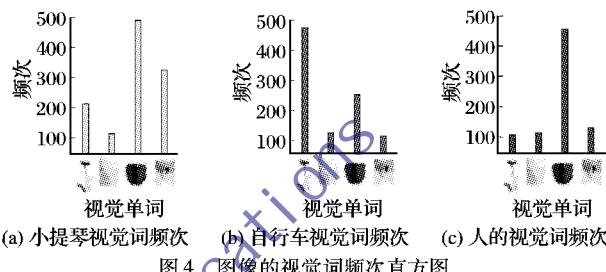


图 4 图像的视觉词频次直方图

1.3.1 PCA 变换

PCA 是设法将原来众多具有一定相关性的 p 个特征,重新组合成一组新的互无关的综合特征来代替原来的特征。通常数学上的处理就是将原来 P 个特征作线性组合,作为新的综合特征。最经典的做法就是用 F_1 (选取的第一个线性组合,即第一个综合特征)的方差来表达,即 $Var(F_1)$ 越大,表示 F_1 包含的信息越多。因此在所有的线性组合中选取的 F_1 应该是方差最大的,故称 F_1 为第一主成分。如果第一主成分不足以代表原来 P 个特征的信息,再考虑选取 F_2 即选第二个线性组合,为了有效地反映原来信息, F_1 已有的信息就不需要再出现在 F_2 中,用数学语言表达就是要求 $Cov(F_1, F_2) = 0$,则称 F_2 为第二主成分,依此类推可以构造出第三、第四,……,第 n 个主成分。通过主成分分析,可以达到对高维特征数据降维的目的。为了后面 Hash 编码的需要,需要利用 PCA 对图像的 500 维特征降到 32 维。

1.3.2 Hash 编码

降维后的特征矢量表示成哈希值其实就是矢量量化的过程,本文采用的矢量量化方式^[5]为:

$$H_{i,k} = \begin{cases} 1, & G_{i,k} > mean_k \\ 0, & G_{i,k} \leq mean_k \end{cases} \quad (3)$$

其中: $H_{i,k}$ 为图像 i 的第 k 维编码, $G_{i,k}$ 是图像 i 的特征向量的第 k 维特征, $mean_k$ 为图像库中所有图像特征的第 k 维均值。因此 K 维的特征值被量化为 K bit,然后把这 K bit 的有序二进制字符串称为这幅图像的 Hash 值。

2 动态距离度量方法

动态距离度量是相对于静态距离度量而言的。静态距离度量是在一个固定的度量空间度量图像特征向量的距离,它只设定一个固定的阈值,该阈值难以准确判断两幅图像是否是近似的。

动态距离度量是根据不同图像的视觉特征动态地自动选取多个度量空间。对于一幅查询图像,首先对该图分别做尺度伸缩、亮度改变、旋转等一系列的变换,变换后的图像被保存到图像库中。接着利用 PCA 方法找到一个空间,在该空间

样本可以被最大限度地区分开。在该空间计算查询图像和各种变换图像的距离，并选取最大的距离作为阈值，在该阈值以内的图像都被保留下来，那些变换后的图像也被保留下来，阈值以外的图像都被过滤掉。在下次迭代中，在保留下来的图像集中寻找一个最具有区分度的投影空间，做相同的操作。迭代过程一直持续下去，直到某一时刻，没有图像被过滤掉，就停止迭代。最后被保留下来的图像就是与查询图像近似的图像。

利用动态距离度量实现近似重复图像检测的步骤如下：

- 1) 对查询图像做多种随机变换；
- 2) 选取区分度最大的特征空间 k ，计算图像的相似度阈值 ε 。 ε 就是查询图像和变换图像之间的最大汉明距离。

$$\varepsilon = \max \left\| \mathbf{P}_{qj} - \mathbf{P}_{q^{(l)}j} \right\|_k \quad (4)$$

其中 $q^{(l)}j$ 为第 l 个随机变换。

- 3) 如果一幅图像和查询图像满足：

$$\sum_{k=1}^L (\mathbf{H}_{i,k} \oplus \mathbf{H}_{j,k}) = 0 \quad (5)$$

和 $\sum_{k=L+1}^K \mathbf{H}_{i,k} \oplus \mathbf{H}_{j,k} \leq \varepsilon$ (6)

则该图像和查询图像是近似的，并把该图像放入下次迭代的数据集 \mathbf{Q} 。即要求前 L bit 具有相同的二进制值，而在其余的 $(K-L)$ bit 中允许有小误差，这里 L 取 24。 \mathbf{P} 是一个投影矩阵， $H(\cdot)$ 为哈希编码函数，距离的计算采用汉明距离。

- 4) 用以下公式更新投影矩阵：

$$\mathbf{P}_i \leftarrow \text{eigenvector}(\text{cov}(\mathbf{Q}), i) \quad (7)$$

$$\mathbf{P} = [\mathbf{P}_0, \mathbf{P}_1, \dots, \mathbf{P}_d] \quad (8)$$

其中： $\text{eigenvector}(\text{cov}(\mathbf{Q}), i)$ 是保留数据集的第 i 位特征向量， d 是低维特征的维度。

5) 重复 2) ~ 4) 直到保留数据集不再改变，最后留在保留数据集中的图像就是和查询图像近似的图像。

3 实验结果

以 1000 个检索词作为关键词（如加菲猫等）搜索网络上的图像，对搜索结果进行人工筛选，同时也对部分图像进行了人工的变形处理，最后形成一个 200 000 张相似图像的检索库。以查准率和查全率两个指标来衡量算法的性能。

选取“加菲猫、天安门、金字塔、奥运标志、鸟巢”五类图像作为查询图像，从图像库中查询近似图像，表 1 是查全率和查准率的实验结果，图 5 是加菲猫的查询结果。

表 1 查全率和查准率的实验结果 %

类别	查全率	查准率	类别	查全率	查准率
加菲猫	83	90	奥运标志	90	93
天安门	80	92	鸟巢	77	92
金字塔	64	95			

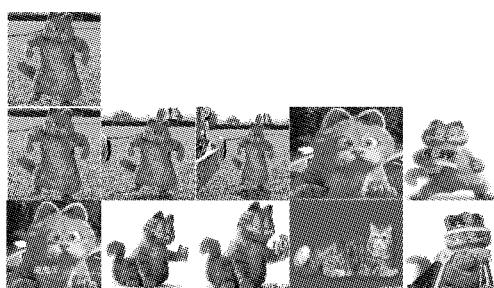


图 5 近似加菲猫图像检测结果

五类图像中，奥运标志的检测结果最好，因为该类图像的背景不复杂，背景对目标的干扰较小。大体上的查准率可以达到 90% 以上，查全率在 70% ~ 80%。

4 结语

本文讨论了近似重复图像检测的常用方法，在此基础提出了基于 Bag-of-words 和哈希编码的高效检测算法。该算法结合了语义级的图像表示、Hash 编码和动态距离度量快速准确检测的特点，和传统的近似图像检测算法相比，本文算法可以动态地计算图像的相似度，可以在查准率和查全率之间做到一个较好的平衡。

基于 k -means 聚类算法形成的视觉单词具有同义性和歧义性的问题，而且具有视觉单词的冗余信息较多，不支持动态扩展等缺点。下一步的工作是利用位置敏感的哈希算法代替 k -means 对 SIFT 特征进行聚类，形成视觉单词。

参考文献：

- [1] SIVIC J, RUSSELL B, EFROS A. Discovering objects and their location in images [C]// Tenth IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2005: 370 ~ 377.
- [2] HAMPAPUR A, BOLLE R M. Comparison of distance measures for video copy detection [C]// IEEE International Conference on Multimedia and Expo. Piscataway, NJ: IEEE Press, 2001: 737 ~ 740.
- [3] ZHANG D Q, CHANG S F. Detecting image near-duplicate by stochastic attributed relational graph matching with learning [C]// Proceedings of the 12th Annual ACM MULTIMEDIA '04. New York: ACM Press, 2004: 877 ~ 884.
- [4] QAMRA A, MENG Y, CHANG E Y. Enhanced perceptual distance functions and indexing for image replicate recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27 (3): 379 ~ 391.
- [5] PHILBIN C J, ISARD M, ZISSERMAN A. Scalable near identical image and shot detection [C]// Proceedings of the 6th ACM International Conference on Image and Video Retrieval. New York: ACM Press, 2007.
- [6] MARET Y, NIKOLOPOULOS S, DUFAUX F, et al. A novel replica detection system using binary classifiers, R-trees, and PCA [C]// International Conference on Image Processing. Piscataway, NJ: IEEE Press, 2006.
- [7] QAMRA L, MENG Y, CHANG E. Enhanced perceptual distance functions and indexing for image replica recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27 (3): 379 ~ 391.
- [8] PHILBIN C J, ZISSERMAN A. Near duplicate image detection: min-Hash and TF-IDF weighting [EB/OL]. [2010-10-10]. http://cmp.felk.cvut.cz/~chum/papers/chum_bmvc08.pdf.
- [9] 蔡博宇, 李娴成, 高毫林. 基于贪婪树的外部支持向量机近似重复图像聚类算法[J]. 信号处理, 2012, 28(4): 601 ~ 606.
- [10] TANG X. Book retrieval based on near-duplicate image matching [C]// Proceedings of the 9th International Conference on Fuzzy Systems and Knowledge Discovery. Piscataway, NJ: IEEE Press, 2012: 2616 ~ 2619.
- [11] WANG G, ZHANG Y, LI F F. Using dependent regions for object categorization in a generative framework [C]// IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2006: 1597 ~ 1604.
- [12] SEMPERE V, ALBERO T, SILVESTRE J. Analysis of communication alternatives in a heterogeneous network for a supervision and control system [J]. Computer Communications, 2006, 29 (8): 1133 ~ 1145.
- [13] 唐坚刚, 王泽兴. 基于 Hash 值的重复图像检测算法 [J]. 计算机工程, 2009, 35(1): 183 ~ 185.