第 35 卷 第 1 期 计 算 机 工 程 2009 年 1 月
Vol.35 No.1 Computer Engineering January 2009
• 图形图像处理 • 文章编号：1000—3428(2009)01—0183—03 文献标识码：A 中图分类号：TP391

基于 Hash 值的重复图像检测算法

唐坚刚，王泽兴
(上海理工大学计算机与电气工程学院，上海 200093)

摘 要：重复图像检测是自动图像标注中经常遇到的问题之一。该文在讨论大规模图像数据库的基础上，提出一种基于 Hash 值的重复图像检测算法。该算法不依赖于具体图像特征，通过建立索引能快速寻找到重复图像，有效提高了查准率。实验结果表明，该算法是可行的，可以应用到其他各种场景。

关键词：图像检测；重复图像；图像内容；Hash 值

Duplicate Image Detection Algorithm Based on Hash Value

TANG Jian-gang, WANG Ze-xing
(Shool of Computer and Electrical Engineering, University of Shanghai for Science & Technology, Shanghai 200093)

【Abstract】 Duplicate image detection is one problem in automatic image tagging. On the basis of discussing the large-scale images database, a duplicate image detection algorithm is proposed, which does not depend on the specific characteristic of images. This algorithm can find the duplicate image by setting up the index, and promote the accuracy of query effectively. Experimental results show the algorithm is feasible and can be used into other relevant cases.

【Key words】 image detection; duplicate image; image content; Hash value

Internet 的发展扩大了图像的传播范围，并提升了其传播速度。人们在享受这种便利的同时，也遇到了由此产生的重复图像问题，即一幅图像会存在多个不同版本。在实际应用中，人们需要检测出这些重复图像。

1 基于 Hash 值的重复图像检测算法

本文描述了一种能在大规模图像数据库中快速而又准确地寻找重复图像的方法，如图 1 所示。

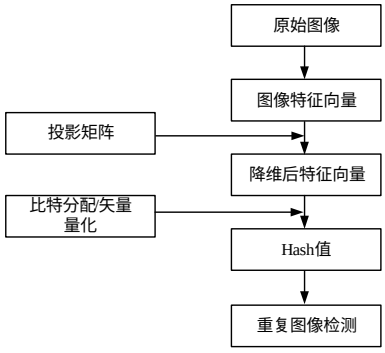

图 1 基于 Hash 的重复图像检测算法流程

对于网络图像搜索引擎而言，以上过程都可在建立图像数据库后完成。利用这些Hash值间的匹配与距离关系，就能进行重复图像检测，由于Hash值的二进制特性，其索引和搜索都能快速[1]地进行。

1.1 图像的特征表示

数字图像本身具有较大的数据量，为便于处理，一般利用从图像中提取出的特征信息来表示图像内容的关键信息。在重复图像检测中，笔者希望这些特征信息能够同时表达颜色信息以及空间结构信息。另外，该特征信息还要有一定的健壮性。当图像经历了图 1 中列举的变化后，其对应特征应当基本稳定。

因此，本文提出采用分块灰度均值(gray block)的方案[2]。如图 2 所示，图像先被均匀地分割为$n\times n$的块。对每一块计算其块内所有像素的平均灰度值，即

$$f_k=\frac{1}{N_k}\sum_{i,j\in B_k} I(i,j)\ ,\quad k\approx 1,2,\cdots,n^2 \tag{1}$$

其中，B_k 代表第 k 个块；N_k 是该块中像素的数目；$I(i,j)$ 是位于坐标 (i,j) 处像素的灰度值。因此，一幅图像可以表示为矢量 $\boldsymbol{F}_i=(f_1,f_2,\cdots,f_{n\times n})^{\mathrm{T}}$。该矢量的维数为 n^2 (一般情况下为高维矢量)。

1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20
21	22	23	24	25

图 2 分块灰度值计算示意图

从另一个角度看，可以把该特征矢量看作是将原图像进行了一次变换，生成一个固定大小为 $n\times n$ 的缩略图像。该特征矢量的计算非常简单，对于每个块只需进行简单的加法和一次除法就能完成，且对图像只需进行一次顺序扫描，有利于对大规模图像数据的高速处理。

基金项目：上海市高等学校青年科学基金资助项目“基于数据挖掘的网络安全管理技术研究”(03SQ05)
作者简介：唐坚刚(1962－)，男，副教授、博士，主研方向：网络信息安全；王泽兴，硕士研究生
收稿日期：2008-08-10 **E-mail：**angel_wzx@126.com

1.2 特征的降维

高维矢量给数据的索引和搜索均带来了一定困难，所以，有必要将其维数降低到一个合理的范围之内。对于重复图像检测而言，特征降维的目标有 2 个：(1)将原来高维的特征降低到一个合理、易于处理的子空间内；(2)由于降维会滤除部分原始信息，因此笔者希望通过这种降维过程来去除一定的噪声信息。这里采用了主成分分析(Principle Component Analysis, PCA)[3]的方法。

$$\boldsymbol{g}_i = \boldsymbol{A}\boldsymbol{f} \tag{2}$$

其中，$\boldsymbol{g}_i$ 为降维后的特征矢量；$\boldsymbol{A}$ 为投影矩阵。

利用PCA进行降维是为了将高维矢量投影到其子空间的投影矩阵，而该矩阵是与实际数据相关的。所以，对图像的遍历要进行 2 次。第 1 次是用来计算协方差矩阵，并求得特征值和与主成分对应的特征向量；第 2 次进行实际的降维操作。当数据量很大时，每次遍历数据库操作都会非常耗时，为节省这些不必要的操作，这里采取了将投影矩阵预先固定的做法，即先构造一个足够大的图像集合，包含约 10^7张来自网络的各种图像，然后利用这些图像得到分块灰度均值特征对应的投影矩阵。对于所有图像，均使用该投影矩阵进行降维。这样，只需对图像数据库进行一次遍历就可得到降维后的结果。

1.3 Hash 值的生成

由降维后的特征向量 $\boldsymbol{g}_i$ 产生Hash值$h_i=\{H_{i,k},\ k=1,\ 2,\cdots,k\}$的过程实际上是种矢量量化过程。如果最终生成的二进制字符串共有K bit且各维之间是独立量化的，那么如何在各维之间分配比特数就是一个很关键的问题。在本文算法中，每维被固定分配 1 bit，且对该位采取了一种简单但很有效的量化方式：

$$H_{i,k} = \begin{cases} 1 & \text{if} \quad G_{i,k} > mean_k \\ 0 & \text{if} \quad G_{i,k} \le mean_k \end{cases} \tag{3}$$

其中，$mean_k$ 代表该维的均值。从而 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 h_i 的值称为这幅图像的 Hash 值。该 Hash 值来自于图像的视觉内容信息，但经过一系列处理后已具有非常简洁的表现形式，有利于进一步索引和搜索。

1.4 重复图像检测

本算法的目标是检测出所有重复图像并构成对应的重复图像组(每个组内可能有多幅图像，每两幅图像互相间均为重复关系)，先将以上操作称为归组(grouping)。在传统 Hash 检索方法中，当给定目标的 Hash 值后，系统在数据库中寻找与其具有相同 Hash 值的数据并将其呈现给用户。由于该搜索过程是基于 Hash 值的精确匹配，因此可使用多种快速索引和存储算法。但可能会受到噪声的影响，造成一定的信息损失。所以，为提高系统性能，不能只寻找与目标图像具有相同 Hash 值的图像，而应寻找与其 Hash 值距离较近的图像。由于 Hash 值是有序的二进制字符串，因此可以定义 Hash 值之间的距离为

$$\text{Hamming}(H_i, H_j) = \sum_{k=j}^{k} (H_{i,k} \oplus H_{j,k}) \tag{4}$$

其中，$\oplus$ 表示二进制加，即异或操作。距离较近意味着两幅图像 Hash 值间的 Hamming 距离小于某一阈值 T。另外，考虑到在 Hash 生成过程中，PCA 可以按照方差由大到小对各维进行排列，在方差较大的那些维上，数据分布较分散，即分界面两面有距离较近点的概率会较小；相应地，方差较小的维上，数据分布较密，相应概率较大。所以，这里规定图像的 Hash 值之间的相似性条件为

$$\sum_{k=i}^{i} (H_{i,k} \oplus H_{j,k}) = 0 \text{ 且 } \sum_{k=i+1}^{k} (H_{i,k} \oplus H_{j,k}) \le T$$

即要求前 L bit(对应方差较大的维)具有相同的二进制值，而在其余的$(K\text{-}L)$ bit 中允许有小误差 T。这一定义不仅能使系统的查全率得到一定提高，而且由于在寻找重复图像过程中，前 L bit 被要求是相同的，因此也给建立快速索引提供了便利。

这里可以通过设置该阈值 T 来调整系统的检测性能。当 T 较小时，图像间的匹配程度较高，准确率就会上升，同时查全率就可能下降。而较大的 T 意味着有较高的查全率，但准确度可能会下降。例如，当该算法被用到版权保护中时，人们往往希望能找到所有可能的非授权图像，因此，这时应采用较高的 T。而在电子商务、图像搜索等应用中，人们不愿受到错误信息的干扰，此时的 T 就可以设得低一些。

2 重复图像检测实验

2.1 数据集

先获取 1 000 个常用的图像搜索查询关键词，通过去除其中的重复词汇和无效词汇，最终得到 995 个有效的查询词汇。将每个词汇提交到网络图像搜索引擎 PicSearch(http://www.picsearch.com)中，并下载其返回的前 1 600 幅图像(每页 16 幅图像，前 100 页)，去除部分无法下载的图像，最终得到 1 443 066 幅图像。

考虑到数据集的大小，在实验中设定 K=24, L=12 且对每幅图像先归类为“自然图像”(Photo)或“计算机图形”(Graphics)。其中，“自然图像”表示该图像来自于一般的成像设备，如自然照片；而“计算机图形”表示该图像是经计算机处理生成的，如动画片中的图像。对于“自然图像”，设 T=1；对“计算机图形”，设 T=0。只有同类图像才能进行比较，以判断是否为重复图像。

2.2 性能度量

由于本算法中引入了图像的归组操作，因此对其性能的评价不宜仅采用传统的查准率和查全率进行衡量。针对重复图像检测，这里提出 4 种性能测度，将它们综合起来就能描述系统的检测性能了。

针对由重复图像构成的“组”提出“组查准率”(Group Precision, GP)以及“组查全率”(Group Recall, GR)。同时可以计算图像对的查准率(Image Pair Precision, IPP)及其查全率(Image Pair Recall, IPR)，即

GP=(正确图像组数目)/(检测到图像组数目)×100%

GR=(正确图像组数目)/(真实图像组数目)×100%

IPP=(正确图像对数目)/(检测到图像对数目)×100%

IPR=(正确图像对数目)/(真实图像对数目)×100%

2.3 重复图像检测性能

表 1 和表 2 分别给出了本文所提算法在手工标注数据上的性能。表中列举的 4 个进行标注的查询词具有一定代表性。其中，“Angelina Jolie”与“Britney Spears”是具有较高关注度的名人，在网络上有大量图像，但这些图像大多来自她们的宣传海报和广告，所以，存有大量重复图像，而且这些图像一般都是“自然图像”；另外 2 个查询词“animal”与“cartoon”都是较抽象的名词，对应的图像内容非常多样化，重复图像较少，而且，这些图像大多是“计算机图形”。

表 1　手工标注数据上的检测性能 1

查询词	检测到的组	正确组	真实组	GP/(%)	GR/(%)
Angelina Jolie	177	176	256	99.4	68.8
Animal	21	21	44	100.0	47.7
Britney Spears	167	164	245	98.2	66.9
Cartoon	59	59	96	100.0	61.5
合计	424	420	641	99.1	65.5

表 2　手工标注数据上的检测性能 2

查询词	检测到的重复图像	检测到的重复图像对	正确的重复图像对	IPP/(%)
Angelina Jolie	276	424	423	99.8
Animal	22	23	23	100.0
Britney Spears	230	327	315	96.3
Cartoon	61	63	63	100.0
合计	589	837	824	98.4

从表 1 中可以看出，本文算法具有较高的组查准率 GP，综合平均值超过了 99%，而组查全率 GR 则相对较低，这是考虑到组查准率 GP 比组查全率 GR 重要而适当调整了系统参数的结果。

即便如此，超过 65%的组查全率 GR 也是能接受的。需要指出的是，在达到同等组查准率 GP 的情况下，“ 计算机图形 ” 的组查全率 GR 较 “ 自然图像 ” 略低。

从表 2 中可以看出，由本文算法检测出的图像对具有较高的准确率，平均达到 98.4%。

表 3 给出了在所有的 995 个查询词上本算法随检测范围变化的性能数据。

从表 3 中可以看出，组查准率 GP 与图像对查准率 IPP 一直保持在 90%以上。因此，本文算法较好地保证了检测出重复图像的准确性。

表 3　不同检索范围的重复图像检测　(%)

	检测范围						
	100	200	500	600	800	900	1 000
GP	95.7	95.0	93.9	93.7	93.3	93.2	93.0
GR	55.4	55.9	56.6	56.7	56.8	56.8	56.6
IPP	96.2	95.4	93.6	93.2	93.1	93.0	92.8
IPR	35.4	32.7	30.2	30.0	29.4	29.3	28.8

3　结束语

本文讨论了在大规模图像集中快速进行重复图像检测以及相似图像搜索的方法，提出一种图像内容的简洁表达方式，每幅图像对应的特征在经过变换后成为一组二进制字符串，即为该图像的 Hash 值。利用该字符串可以快速而准确地进行重复图像检测。实验结果表明，该方法计算复杂度低，准确度高，且所需的附加存储量小，具有很好的性能。

但由于图像往往被表示为高维空间中的矢量，因此在现有的技术条件下对其索引和检索都是较困难的，而且 “ 语义鸿沟 ” 问题同样影响了图像检索结果。今后的工作将通过机器学习的方法对这些信息进行组合，以期更加快速准确地找出其重复或相似图像。

参考文献

[1] 曹　炬, 马　杰, 谭毅华, 等. 基于像素抽样的快速互相关图像匹配算法[J]. 宇航学报, 2004, 25(2): 173-178.

[2] 王亮申, 欧宗瑛. 图像纹理分析的灰度一基元共生矩阵法[J]. 计算机工程, 2004, 30(23): 78-80.

[3] 覃冬梅. 一种基于主分量分析的恒星光谱快速分类法[J]. 光谱学与光谱分析, 2003, 23(1): l82-186.

（上接第 158 页）

地址(srcmac)为 AP 的 MAC 地址，表示对 AP 身份的冒充。

4)$t7$，测试人员向合法用户发送伪造的连接断开命令使其断开连接。

5)$t9$，当被断开的合法用户重新与 AP 连接时，测试人员对其连接过程进行窃听，获取连接时需要传输的 SSID。

针对 MAC 地址过滤的攻击测试包括：

1)$t4$，获取某个合法用户的 MAC 地址。

2)$t8$，当获得某个合法用户的 MAC 地址，且该用户已经与 AP 断开连接后，测试人员就能冒充该用户接入 WLAN。

针对 WEP 攻击测试包括：

1)$t2$，窃听无线网络数据并将其储存。

2)$t5$，由于WEP使用的初始化向量长度仅为 24 bit，因此当收集消息的数量n大于阈值$TH(TH=2^{24})$时，可以对已收集的消息进行分析从而获得WEP密钥。

假设变迁t_1-t_9对应的权值分别为 1,1,2,2,6,2,1,2,1，则　3 种测试用例按其权值由小到大依次排列为 t_1,t_3,t_6,t_7,t_9; t_1,t_4,t_8; t_2,t_5。按测试路径对测试目标进行攻击测试并记录测试结果，测试过程中可以使用各种脚本与安全测试工具辅助测试。测试结果表明上述 3 种针对WLAN的渗透测试可以达到测试目标。

(5)漏洞泛化。测试中确认的漏洞主要体现在身份认证机制和密钥管理机制上，测试目标仅由 AP 用 MAC 地址对用户进行单向身份认证。WEP 中 IV 的空间很小，容易出现重复使用密钥的情况，应检查系统中使用身份认证与数据加密机制的其他部分，以确证是否存在相似漏洞。

(6)漏洞消除。根据确认的漏洞，可以采取以下手段来增强 WLAN 的安全性：采用客户端和 AP 的双向认证，只有双方相互认证完成后才可以访问网络；动态改变 WEP 的密钥；使用具有更高安全性的 802.11i 进行认证与数据加密等。

3　结束语

Petri 网模型对离散事件具有强大的描述能力且直观可靠，本文将其应用到渗透测试过程，给出不同攻击场景的合成规则和测试用例生成算法。对无线局域网的渗透测试实例证明了 Petri 网在攻击模拟以及测试过程组织方面的优势。如何把 Petri 网的层次特性应用于对复杂攻击过程的建模方法、在复杂场景中测试用例生成效率的问题以及如何利用建立攻击测试网模型生成自动化测试工具等有待进一步研究。

参考文献

[1] 张继业, 谢小权. 基于攻击图的渗透测试模型的设计[J]. 计算机工程与设计, 2005, 26(6): 1516-1518.

[2] McDermott P. Attack Net Penetration Testing[C]//Proceedings of the 2000 Workshop on New Security Paradigms. New York, USA: ACM Press, 2000.

[3] 杨义先, 钮心忻. 无线通信安全技术[M]. 北京: 北京邮电大学出版社, 2005.