

# 在 World Wide Web 上检索图像和视频信息

朱磊

(解放军理工大学通信工程学院, 210007, 南京)

[摘要] 提出了一种综合利用图像和视频的文本信息和基于内容的视觉特征进行 World Wide Web 上图像视频信息检索的原型系统. 在这个原型系统中, 一个完整的 Web 上图像和视频信息的处理过程包括: (1) 通过探索器从 Web 上自动收集信息; (2) 在文本和视频特征两个域内同时进行分析; (3) 图像和视频信息分类; (4) 标定索引, 进行快速检索. 实验结果表明, 利用这个系统, 可以获得较高的视觉分类率.

[关键词] 基于文本内容的图像处理, 基于图像内容的图像处理, 图像分类, 基于内容的检索

[中图分类号] TN911. 73, [文献标识码] B, [文章编号] 1672- 1292- (2004) 01- 0046- 04

## 1 简介

Web 上的视频、图像信息是处于经常的变化之中的, 对 Web 上视频和图像信息的分类是通过高效的自动系统来实现的, 该系统定期从 Web 搜寻图像和视频信息, 并向用户提供一种便捷的检索和获取的操作方式. 本文提出了一种新的对 Web 上的图像和视频的基于内容的检索系统, 其创新点在于综合了基于文本内容的检索方式和基于图像视觉特征的检索方式对图像和视频信息进行分类和处理.

图像和视频的主题和类型信息可同时从文本信息和视觉信息中获得, 文本信息可以是 Web 地址或上一级文档的参考文本等. 目前的知识分类学对于视觉信息来说不是很适合, 因此, 本文提出了一种新的视觉信息的分类处理方法. 新的分类方法是建立在图表结构基础之上的, 图像和视频信息通过一些全自动或半自动的处理过程被分类, 分类的第一步, 使用关键术语字典, 根据从图像、视频相关文本中检测到的关键术语, 对其进行分类. 第二步, 从图像、视频的 Web 地址中提取关键术语, 进行人工分类.

基于图像内容的分类和处理技术提供了对一些显著的视频特征如颜色、纹理、轮廓、空间位置等的自动评估. 通过计算图像间这些提取出来的特征的相关性, 可以实现根据图像的特征数据进行查询、自动对图像和视频进行同类场景的组合、通过图像或视频文件进行浏览和导航等功能. 基于内容的检索工具可通过学习达到提高检索效果的目的. 相关反馈就是一种学习模式, 它的一种形式是用户从检索返回的结果集中再选择若干项作为下一次

检索的条件. 它的另一种形式是用户从目前的结果集中选择最接近或最不接近所需的图像或视频作为新的检索条件, 通过相关反馈技术, 系统可以自动重新形成检索条件以更好地匹配用户的需求.

## 2 图像和视频的收集过程

图像、视频的收集过程是由若干个自治的 Web 代理或探索器实现的. 它们根据文档之间的超级链接关系, 在 Web 上漫游, 发现图像、视频信息进行下载, 并处理这些文档, 将新的信息加入到分类目录中.

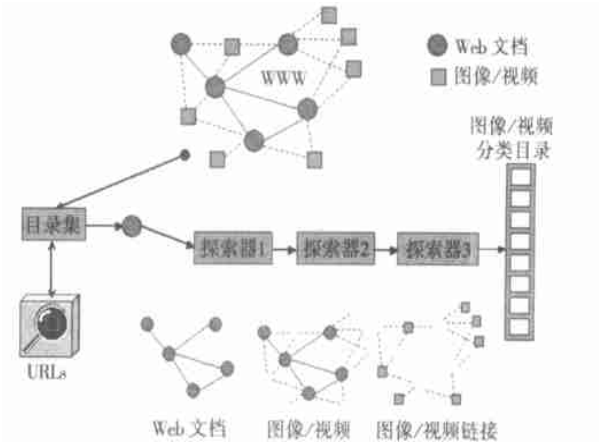


图 1 图像和视频信息通过探索器在 Web 上的采集过程

整个过程是由几个独立的探索器来实现的: 探索器 1 负责收集有可能包含图像、视频的 Web 页或有相关超链接的文档; 探索器 2 负责解析图像、视频页的 URL; 探索器 3 负责获取并分析这些图像和视频.

图像和视频信息的检索由两个阶段组成. 如图 2 所示, 检测的第一阶段由 2 个探索器在网络上巡

游,寻找图像和视频信息.探索器 1 从初始 URL 出发,在 Web 上进行宽范围的搜索,通过 HTTP 协议下载网页,并将 HTML 代码交付给探索器 2.探索器 2 从 HTML 中检测出新的 URL,将其加入到探索器 1 等待下载的队列中.探索器 2 对网页上的所有超链接进行检测,把相关的 URL 转换成绝对地址.根据检查超链接的类型和 URL 中文件的扩展名,探索器 2 将 Web 文档分成几种类型:图像、视频和普通 HTML.第二阶段,探索器 2 中列出的包含图像和视频信息的 URL 被交付给探索器 3,探索器 3 获取这些图像和视频,对它们进行处理,并根据处理结果将其加入到分类表中,如图 3 所示.探索器 3 的主要功能有:①提取出基于内容的视觉特征,如颜色直方图等;②提取出其他的属性信息,如宽度、长度、画面个数、视频数据的类型等;③生成足够表征视觉信息的缩略图.

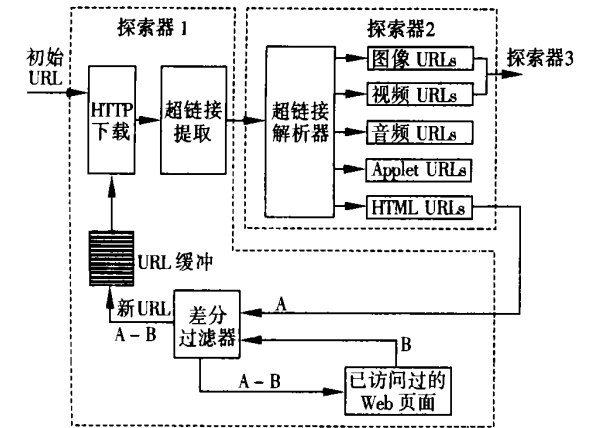


图 2 探索器 1 和探索器 2 在 Web 上巡游并生成图像和视频的 URL 序列

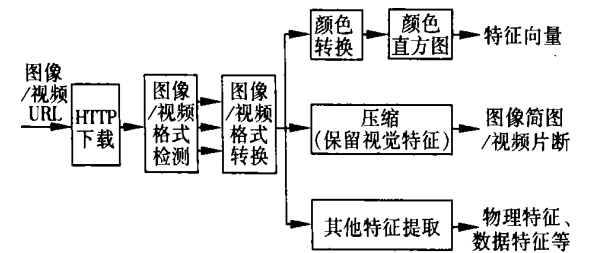


图 3 探索器 3 对图像和视频信息的处理过程

3 基于文本内容的图像和视频信息处理

在图像和视频信息的自动分类中可以考虑一些伴随的文本信息.在 Web 上,每一个图像和视频都有唯一 Web 地址,有的还可包含其他的 HTML 标签,这些都为说明图像和视频的内容提供了信息.通过对图像和视频的 Web 地址、URL 和 HTML 标签等文本信息进行处理,可以对图像和视频标定

索引.

对文本信息的处理包括术语提取、目录名称提取、用关键术语字典实现从关键术语到主题自动映射和半自动的从目录名称到主题的映射.

图像和视频在 Web 有两种发布方式:联机和引用.这两种情况下的 HTML 语法是有区别的.对于联机方式或嵌入方式,在 HTML 文件包含以下的代码:<img src= URL alt= [alt text]>,其中,URL 给出了图像或视频的相对或绝对地址.可选的 alt 标签指定了当浏览器加载图像或视频时可能出现的代替图像或视频的文字,或当浏览器找不到图像或视频时出现的代替文字.另外,图像和视频也可被上一级页面引用,引用的 HTML 语法为:<a href= URL> [hyperlink text] </a>,其中,可选的[ hyperlink text]代表超链接所指向的图像或视频对象.通过对 URL、alt 标签和 hyperlink text 进行非字母字符分割,从图像、视频中提取出术语.通常,图像或视频的 URL 具有如下的形式:URL= http://host. site. domain[: port]/[ user/][ directory/][ file[. extension]],其中,[...]中的内容是可选的.例如,有下面 3 个 URL:

URL<sub>1</sub> = http://www. mynet. net: 80/ animals/ domestic- beasts/ dog37. jpg

URL<sub>2</sub> = http://Camille. gsfc. nasa. gov/rsd/ movies2/ Shuttle. gif

URL<sub>3</sub> = http://www. arch. Columbia. edu/DDI/ projects/ amiens/ slides/ slide6b. gif

从 directory 和 file 串中提取出术语,并用 F<sub>key</sub>和 F<sub>chop</sub>表示:

$$F_{key}(URL) = F_{chop}(directory/file) \tag{1}$$

其中, F<sub>chop</sub>( string)列出了由非字母字符分隔的子字符串,例如, F<sub>key</sub>(URL<sub>1</sub>) = F<sub>chop</sub>(" animals/ domestic- beasts/ dog37") = " animals", " domestic", " beasts", " dog"术语可以通过字符串的匹配来实现基于文本的查找.在提取出术语后,系统就根据变换后的文本对图像和视频进行索引.

从 URL 中提取出的目录名称说明了图像和视频在 Web 上的位置信息,它包含了 URL 的目录部分,即 F<sub>dir</sub>(URL) = directory,目录名同样可以在图像、视频和主题分类之间进行映射.

4 基于图像内容的图像和视频信息处理

本文利用色彩直方图对图像和视频信息的内容进行描述.目前的研究结果表明,特定的基于内

容的特征只适合于特定的应用领域,而色彩直方图描述了图像颜色的统计分布特征,且具有平移、尺度和旋转的不变性,使用直方图可以实现与域无关的方法,因此在颜色检索中被广泛采用。

色彩直方图描述了在图像和视频中的颜色分布状况。系统在量化的 HSV 色彩空间中计算每一个图像和视频场景的色彩直方图,以评估它和其他图像或视频场景之间的相关度。色彩直方图还被用来为图像和视频自动分类,分类是利用 Fisher 判别式实现的。

#### 4.1 直方图相似性度量

本文中,采用直方图相异函数来度量直方图之间经过加权的相异程度。查询条件的直方图  $h_q$  和目标直方图  $h_t$  之间的二次项距离由下式给出:

$$d_{q,t} = (h_q - h_t)^T A (h_q - h_t). \quad (2)$$

其中,  $h_q$  和  $h_t$  为列向量,  $A = [a_{i,j}]$  是一个对称矩阵,  $a_{i,j}$  表示了色彩  $i$  和  $j$  之间的相关度,有  $a_{i,i} = 1$ 。

$h$  为归一化直方图,有  $\|h\| = \sqrt{\sum_{m=0}^{M-1} h[m]^2} = 1$ 。

为了提高在色彩直方图查询过程中的有效性,对色彩直方图的二次方程进行分解,这将能提高计算和索引的效率。定义  $\mu_q = h_q^T A h_q$ ,  $\mu_t = h_t^T A h_t$  以及  $r_t = A h_t$ , 得出色彩直方图的二次项距离公式为:

$$d_{q,t} = \mu_q + \mu_t - 2 h_q^T r_t \quad (3)$$

将矢量  $r_t$  分解成元素形式:  $r_t[m]$ , 距离公式可以通过设置参数  $\tau$  近似地获得任意的精确度:

$$d_{q,t} - \mu_q = \mu_t - 2 \sum_{\forall m \text{ where } h_q^T[m] r_t[m] \geq \tau} h_q^T[m] r_t[m] \quad (4)$$

这样,对  $h_q$  的最相关色彩直方图的查询处理过程可以简化为索引单个的  $\mu_t$  和  $r_t[m]$ 。对于查询来说,  $\mu_q$  是个常量,与之在色彩直方图上最接近的  $h_t$  就是使得  $\mu_t - 2 \sum_{\forall m \text{ where } h_q^T[m] r_t[m] \geq \tau} h_q^T[m] r_t[m]$  取最小值的。

#### 4.2 自动类型评估

在图像和视频的色彩直方图的基础上,本文提出了一个用 Fisher 判别式进行自动类型评估的处理过程。Fisher 判别式为色彩直方图构造了一系列相互无关的线性加权值,用来为训练类别之间提供最大的防卫度。由于这些线性加权值由第  $K$  类图像的距离特征向量矩阵(同一类内向量距离和类之间的向量距离)决定,因此,可以从色彩直方图之间的距离来判定图像和视频所属的类别。对一个

新的色彩直方图  $h_n$ , 根据下式自动判别与它最接近的类别为  $K$ :

$$[T(h_n - m_k)]^2 \leq [T(h_n - m_i)]^2, \forall i \neq k \quad (5)$$

其中  $T$  是从训练的类中提取出的特征向量矩阵,  $m_i$  是  $i$  类型的色彩直方图。

#### 4.3 相关反馈

用户通过决定返回集中哪些是和查找条件相关的,哪些是无关的,来更好地实现检索。利用色彩直方图,可以通过以下方法实现相关反馈:令  $I_r = \{\text{相关的图像和视频集}\}$ ,  $I_n = \{\text{无关的图像和视频集}\}$  是由用户决定的。在  $K$  次循环反馈后,第  $K+1$  次循环的查询矢量  $h_q^{k+1} = \left\| \alpha h_q^k + \beta \sum_{i \in I_r} h_i - \gamma \sum_{j \in I_n} h_j \right\|$ 。其中,  $\|\cdot\|$  表示归一化处理。这样,通过  $h_q^{k+1}$  和公式(4)的计算,就可以获得新的返回图像。可将上式简化,令  $\alpha = 0$ ,  $\beta = \gamma = 1$ , 这种情况下,给正向的图像和反向的图像赋予相同的加权值。

### 5 图像和视频的主题分类和检索

结合文本内容信息和图像内容信息,利用建立关键术语字典,对图像和视频进行主题分类,主题分类表达了图像和视频的语义内容。关键术语是经过人工标识的,与主题相关联的术语项。关键术语字典包括了关键术语集和与它们相关的主题分类。本文中采用半自动的方式建立了关键术语字典。用户的检索条件可以是文本的和图像的,首先根据文本内容信息进行粗分类,然后计算图像和视频的术语直方图,然后按照各术语出现的频次倒序排列,以提供给系统进行分类评估。

对图像和视频信息的检索流程如图4所示。图中,  $A$  列表是查找的结果列表,  $C$  列表是用来进行反馈操作的列表,用户通过增加或减少记录对  $C$  进行操作。用户通过对  $A$  和  $C$  进行不同逻辑关系的运算,得出一个结果  $B$ 。  $A$  和  $C$  之间的逻辑操作可以是以下几种方式之一:

$$B = A \cup C, B = A \cap C, B = C - A, B = A, B = C.$$

对得到的结果集  $B$ , 用户可以基于内容的或是基于文本的方式进行浏览和查询。浏览和查询的结果输出集是  $C$ ,  $C$  集是  $B$  集经过排序后的子集,  $C$  集的排序是按照和用户所选项目的相关度排定的。这种浏览和查询可以是针对输出集  $B$  或是整个分类集合的,在前一种情况下,用户浏览当前的输出结果集,并从中选择某个项目,然后系统返回与用

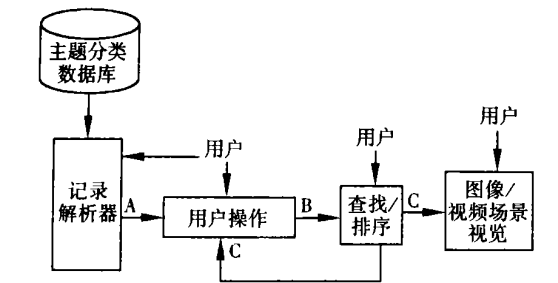


图 4 检索过程

户选定的项目相关性最高的记录。

6 评估

在系统的实验中, 对分布在 16 773 个独立的 Web 站点上的 46 551 个目录下的 513 323 个图像和视频进行了分类, 分类如表 1 所示。

表 1 实验结果数据

项目	统计值
被分类的图像和视频	513 323
提供图像和视频的站点	16 773
独立的 Web 目录	46 551
黑白或灰度图像比例	14. 15%
视频信息比例	1. 05%
被分到相应主题类别的图像和视频比例	68. 23%
主题分类大小(类别数)	941
关键词语字典大小(术语数)	932
目录名称到主题映射数	1 018

从上面的表中可以看出, 通过使用关键词语的自动映射和目录名称的半自动映射, 有 68. 23% 的图像和视频实现了主题分类。对从 941 个主题类中随即选取的几个主题类进行评估, 得出的结论是整体的主题分类性能较好, 达到了接近 92% 的精确度。视觉分类评估的分类效果很好, 高达 95%。

同时我们发现在下列的集中情况下会发生分类错误: (1) 关键词语的使用超过了图像、视频发布

的上下文理解(URL 中提取出的术语不是图像、视频实际表达的内容); (2) 分类所依赖的关键术语具有多义性; (3) 分类所依赖的关键术语出现在目录名称中。在进一步的工作中, 我们将引入其它的视觉特征, 如纹理、形状轮廓和空间位置层次关系等对图像进行描述以提高基于内容的理解。另外, 将建立关键词语字典的过程由半自动改进为全自动, 并进一步对分类过程的自动化处理方式进行研究。

[ 参考文献]

[1] Gudivada V N, Raghavan V V, Grosky, W I, et al. Information retrieval on the World Wide Web[J]. IEEE Internet Computing, 1997, 1( 5): 58~ 68.

[2] Jung G S, Gudivada V N. Autonomous tools for information discovery in the world-wide web[ D]. School of Electrical Engineering and Computer Science, Ohio University, Athens, OH, 1995.

[3] D Zhong, Zhang H J, Chang S F. Clustering methods for video browsing and annotation[ D]. In Symposium on Electronic Imaging: Science and Technology-Storage & Retrieval for Image and Video Databases IV, volume 2670, San Jose, CA, February 1996. IS&T/ SPIE.

[4] Guglielmo E J, Rowe N C. Natural language retrieval of images based on descriptive captions[ J]. ACM Trans Info Systems, 1996, 14: 237~ 267.

[5] Smith J R, Chang S F. Querying by color regions using the VisualSEEK content-based visual query system[ A]. Maybury M T. Intelligent Multimedia Information Retrieval[ C]. IF-CAI, 1996.

[6] Rocchio Jr J J. Relevance feedback in information retrieval: Gerard Salton[ A]. The SMART Retrieval System: Experiments in Automatic Document Processing[ C], New Jersey: Prentice-Hall, Englewood Cliffs, 1971. 313~ 323.

[7] Guglielmo E J, Rowe N C. Natural language retrieval of images based on descriptive captions[ J]. ACM Trans Info Systems, 1996. 14: 237 ~ 267.

Retrieval of Images and Videos on the World Wide Web

Zhu Lei

(Institute of Communication Engineering, PLAUST, Nanjing 210007, PRC)

**Abstract:** In this paper, a prototypical visual information system for searching for images and videos on the World Wide Web is proposed. In the prototype, images and videos are catalogued by combining text-based processing with content-based visual analysis of the images and videos. Besides, a complete processing procedure for images and videos on the web has been studied, including (1) information collecting by automated agents, (2) processing in both text and visual feature domains, (3) cataloging image and video, (4) making index for fast search and retrieval. The experimental result shows that a higher cataloging performance can be achieved by using the prototype.

**Key words:** text-based image process, content-based image process, image classification, content-based retrieval

[ 责任编辑: 刘健]