中华文化基因在线服务平台①

赵海英[1], 彭 宏[2], 陈 洪[3]

[1](北京邮电大学世纪学院 移动媒体与文化计算北京市重点实验室, 北京 102101)
[2](新疆师范大学 网络教育学院, 乌鲁木齐 830054)
[3](中国农业大学 信息与电气工程学院, 北京 100083)

摘 要: 中华文化基因是传承民族文化的基本信息单元, 对它的提取有利于中国传统文化的保护及传承. 但中华文化内涵的深度挖掘以及文化演变规律的发现都是极具挑战性. 文化基因是一个文化模式与文化特征的表征单位, 这是人类文化内涵的密码组. 通过文化基因的表征可以深度挖掘文化内涵, 为人类文化的释义提供深层理解和解析. 论文基于文化基因的标识, 实现文化基因上传、下载; 图像、语音检索等功能, 提供中华文化基因在线服务, 并基于内容聚合的爬虫技术, 丰富中华文化资料库, 传承中华民族文化精髓.

关键词: 文化模式; 文化特征; 文化基因; 在线平台; 网络爬虫; 直方图

Online Platform of Chinese Culture Genes

ZHAO Hai-Ying[1], PENG Hong[2], CHEN Hong[3]

[1](Digital Culture and New Media Technology Research Center, Century College, Beijing University of Posts and Telecommunications, Beijing 102613, China)
[2](College of Network Education, Xinjiang Normal University, Urumqi 830054, China)
[3](School of Automation, University of Science and Technology, Beijing 100083, China)

Abstract: The extraction of the Chinese culture gene is of greatly importance to the protection and development of our country's culture. But the deep excavation of cultural heritage and the discovery of evolution connotation are challenging. Cultural gene is a characterization unit of cultural patterns and cultural features, a human connotation password set. Through representation of culture genes, we can deeply mine culture in a bid to provide a deeper understanding to the interpretation of human culture. Based on the identity of cultural genes, it provides the cultural gene upload, download, search, voice recognition and other functions. The platform also completes the crawler-based content aggregation technology to enrich the cultural heritage database, and inherit Chinese culture.

Key words: equilateral pattern primitive; pattern tile; triangle pile; random pattern

1 引言

5000 年的文明使中国蕴藏了不可估量的文化遗产及非物质文化遗产. 仅据现在统计, 截止 2014 年中国就有 45 项世界遗产被联合国教科文组织批准列入《世界遗产名录》, 位居世界第三; 非物质文化遗产达到了 38 项, 位列世界第一. 然而在这些被世界公认的遗产背后, 还有成千上万的文物在历史的长河中被埋没、被遗忘. 一旦这些文物失去联系, 无法继承 5000 年的文化底蕴, 这将对中国历史的探索、对中华文化的研究产生巨大的影响. 因此, 探索构建一个文化遗产数字化体系势在必行, 文化遗产数字化将会把这些资料、文献和文物等一并纳入囊中, 以供后人研究与查阅[1].

目前, 国内外已有一批较为成熟和成功的数字化保护成果. 但在研究媒体技术数字化文化和文化内涵挖掘的同时, 深入研究文化基本元素提取, 而不仅仅

① 基金项目:国家自然科学基金(61163044);国家社科基金重点项目(12AD118-2,12AZD120-2);北京市科委项目(Z141110004414074,Z141100001914035)
收稿时间:2015-03-25;收到修改稿时间:2015-05-23

是文物或遗产本身. 文化基因即是民族文化在历史长河中传承、延续所依靠的就是具有内在联系的、富有生命力的核心元素以及遗传密码[2]. 通俗易懂的来说文化基因就是如 DNA, 如同传承人类生命和特征一样传承着文化, 例如儒家思想的精髓、文字的组成、琴曲的音调基因, 又或是组成一幅地毯的图案基因[3]. 对于文化基因的研究相比于单纯对文物或遗产的记录更具有深远的意义.

本文探讨实现中华民族文化基因在线服务平台的关键技术. 包括: 基于主题的聚焦网络爬虫技术, 抓取网络与文化遗产相关的主题内容; 基于颜色和纹理特征的图像内容检索方法和基于文化基因的图案再设计, 提取织物图案基因, 应用于基于文化内涵的图案再设计, 提升民族文化传承方式等.

2 相关工作

国外在文化资源的数字化研究最早可追溯至 R. Fielding 于 1965 年首次出版的《The technique of special-effects cinematography》, 该书主要阐释了如何在电影、动画电影中创作专业视觉效果等问题, 在数字化时代, 这些传统的特效技巧成为后世文化资源数字化展示与传播、虚拟体验等的重要借鉴[4]. J. Gomez-Lahoz 和 D. Gonzalez-Aguilera(2009)设计实现了一种低成本的、灵活的、自动化水平较高的数字化系统, 用于虚拟考古遗址的建模[5]. H. Balk 和 L. Ploeger(2009)的研究, 首次解决了许多图书馆所面临的大批量历史印刷资料扫描件的全特征电子文本转换问题, 并在最少人工干预的前提下显著改善了用户的可访问性[6]. 20 世纪 90 年代以来, 我国数字文化遗产主要是对古代遗址、古建筑、石窟、壁画、雕塑以及文学为代表的文化遗产和音乐、戏曲、民族民间习俗和其它非物质文化遗产进行数字化复制和挖掘, 取得了重要成果, 同时国内许多学者将当前流行的数字化技术引入中华文化资源上进行了诸多尝试和总结, 提出了一系列创新性、改进性的新技术与算法. 其中路遥等(2005)提出了一套虚拟旅游系统体系, 着重分析了人机界面、网络虚拟现实等关键技术[7]. 杨安祺和殷耀文(2006)利用虚拟现实的建模语言 VRML 在 VrmlPad 上建立了一个类似于博物馆的全景 3D 模型, 把真实的秦郑国渠虚拟成计算机中的三维世界, 并实现了虚拟漫游和网络互动[8].

综上所述, 国内学者在文化遗产数字化技术上进行了诸多研究, 并取得了一系列成果. 然而, 现有的文化遗产数字化保护技术还仅仅停留在对于原始风貌影音记录. 近年来, 随着学者们对中华文化内涵的深入研究, 均认可中华文化是由固有文化基因组成的. 因此, 如果能将文化分析的粒度细化到文化基因层次, 则不仅可以起到保护文化的目的, 还可以通过分析文化基因的特性, 基因间的关系挖掘中华文化的真正精髓, 掌握中华文化的发展演变, 以此指导对民族文化内涵揭示, 单个、片断文化遗产的整合, 再现其历史文化原貌等诸方面, 有着突破传统传承和保护的新认识.

3 中华文化基因在线服务平台设计

中华文化基因在线服务平台主要由两大模块组成: 数字化和检索系统, 其中数字化系统的功能是将基因收集入库. 其中两种方式, 一是手动上传, 即将一些本来只存在于博物馆中的文物、字画进行以照片、录影等数字方式进行储存, 并上传到系统中. 二是通过网络爬虫技术进行数据采集, 这种方式可大大提高数据的收集效率. 检索系统分为文本检索、基于内容的图像检索、基于语音的文本检索及基于地理位置的检索. 平台总体结构如图 1 所示.

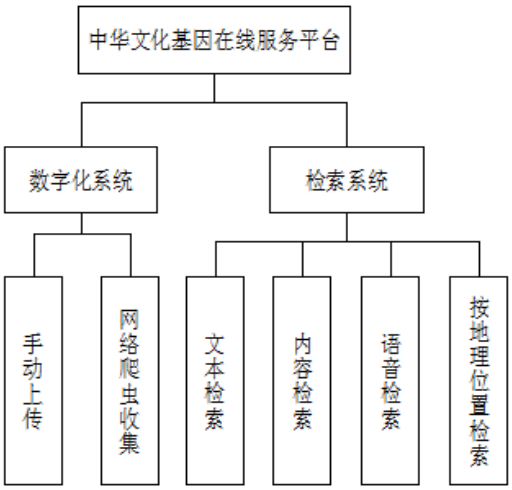

图 1 中华文化基因在线服务平台模块图

3.1 基于爬虫技术的文化基因库内容聚合

相对传统爬虫, 聚焦爬虫添加了链接评价模块及内容评估模块, 即过滤模块.

(1)页面获取模块: 该模块通过给定的主题相关的 URL 进入对应的页面. 它的性能直接影响网络爬虫的

工作性能, 因此需要特别注意多线程、超时以及异步设计.

(2)页面分析模块: 该模块对获取的页面进行解析, 提取页面中存在的链接, 同时保存相应的页面; 其中链接提取功能对于搜索引擎的查全率有一定的影响.

(3)链接过滤模块: 该模块主要是过滤重复下载链接.

(4)链接队列模块: 该模块用来保存和管理链接, 采用队列结构, 以方便提供给页面获取模块使用.

(5)链接评估模块: 主要目的是将重要性较高的链接放入优先访问列表. 列表的排序则是通过对链接进行的重要性评估.

(6)内容评估模块: 内容评价模块主要目的是通过对网页内容分析, 来预测网页中链接的重要性. 该模块是通过级联分类器或 SVM 依据主题训练出一个分类模型, 然后根据网页内容与模型的相似度进行计算, 当达到 70%或者更高的时候则认为该页面为重要页面. 依此进行页面链接重要性评估.

3.2 基于 CBIR 的文化基因库检索

本模块算法流程如图 2 所示:

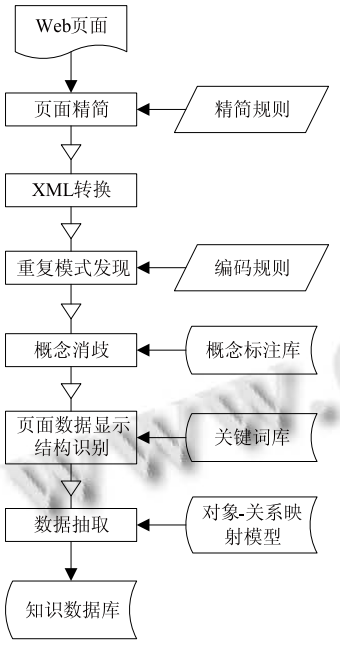

图 2　Web 信息抽取流程

CBIR 是基于传统的文本检索无法描述多媒体内容而产生的, 是通过分析图像的底层或语义特征以及上下文关联以达到基于内容的图像检索功能. 在特征提取方面, 颜色特征是图像不可或缺的视觉特征之一. Swain 等人[10]提出的颜色直方图、颜色相关图、 颜色矩法[11]及颜色一致性矢量等作为颜色特征, 但是这些方法缺少颜色与图像的空间信息的关联, 为此, 最简单的方法是将图像分块, 例如分成九块大小相等的子图像, 继而统计每个子图像中的颜色直方图[12], 建立了颜色与空间关系; 为此 Pass[13]等又提出了颜色聚合向量方法. 在纹理特征方面, Haralick 等人[14]提出了灰度共生矩阵描述纹理特征. 该方法以反差、能量、熵、相关等统计量作为特征量, 较好地应用到一些商业图像检索平台中. Tamura 等[15]人提出了另外不同方法来描述纹理特征, 将纹理特征描述为光滑、粗糙、粒状等特征. 另外, 形状特征也是刻画物体本质特征之一, 同样是描述图像内容的一个重要特征. 早期jain 等人[16]使用封闭直线段来描述形状, 庄挺越提出的内角直方图概念等都存在一定的局限性, 即描述形状不能独立于形状的方向、大小和位置.

本文先将从后台获得的图像数据由 RGB 空间转为 HSV 空间, 其原理是依据人的视觉分辨能力及色彩的不同范围进行非等值量化, 将 HSV 三个分量分成 H 空间划分 16 个等级、S 空间划分 8 个等级, V 空间划分 8 个等级, 并分配各自的变化范围.

纹理特征是使用灰度共生矩阵计算纹理特征量, 首先初始化矩阵. 然后在水平、垂直及对角方向进行计算. 最后计算 5 种特征值: 熵(ENT)、能量(ENE)、对比度(CON)、自相关(COR)和同质性(HOM). 设 G(大小为 k*k)为灰度共生矩阵, 则这 5 种特征值的计算公式为:

$$ENT = -\sum_{i=1}^{k}\sum_{j=1}^{k} G(i,j)\log G(i,j)$$

$$ENE = \sum_{i=1}^{k}\sum_{j=1}^{k} (G(i,j))^2$$

$$CON = \sum_{i=1}^{k}\sum_{j=1}^{k} (i-j)^2 G(i,j)$$

$$COR = \sum_{i=1}^{k}\sum_{j=1}^{k} \frac{ijG(i,j)-u_i u_j}{s_i s_j}$$

$$HOM = \sum_{i=1}^{k}\sum_{j=1}^{k} \frac{G(i,j)}{1+|i-j|}$$

3.3 文本检索技术

考虑到基于内容的图案检索存在语义鸿沟, 系统也提供了文本检索功能. 对一幅图案中基因进行文本检索, 减小图案基因在文化内涵之间领域专家与底层

特征之间的代沟.

文本检索是指对一段或一组自由文字记录的用户匹配, 这些文本可以是任何类型的非结构化文本. 从技术层面来说, 基于文字的检索主要根据文档的文字内容来计算查询和文档的相似度, 最终返回一个匹配值及相应排序结果. 目前存在三种经典模型: 布尔模型、向量空间模型和概率模型. 其中, 在布尔模型中, 一个查询词包括了关键词以及逻辑运算符, 总称为布尔表达式. 它只设定 0 假、1 真两种相关度, 并仅返回为 1 真的文档. 形式十分清楚简单, 且返回内容精准. 向量空间模型相比于布尔模型要求的准确匹配, Salton 在 60 年代末提出的向量空间模型(SVM)采用了“部分匹配”的检索策略. 该模型中用户的查询和信息都表示成关键词及其权重构成的向量. 通过计算向量之间的相似度可以将与用户查询最相关的信息返回给用户. 概率模型, 主要在 1976 年由 Roberston 和 Sparck Jones 提出的二元独立概率模型. 它是通过概率方法将查询和文档关联起来, 通过估计文档与查询相关联概率, 并依据关联概率对所有文档进行排序.

3.4 文化基因库设计

文化基因库数据共分为 5 部分, 分别为城市表(CITY: 图 2)、博物馆表(LIBRARY: 图 3)、馆藏表(ARTICLE: 图 4)、特征库表(FEATCHER: 图 5)、错误记录表(ERROR: 图 6), 其中,

城市表主要用于存放全国的 365 座城市的 id.

博物馆表主要用户存放爬虫爬取得基于 365 座城市的博物馆.

馆藏表主要是存放爬虫爬取的博物馆中的文物列表.

特征库表主要存放图像检索管理员批量入库时的图像特征存放, 以供图像检索使用.

错误日志表主要存放爬虫数据获取失败时的错误记录. 提供依据供爬虫重新爬取.

#	Name	Datatype	Length/Set	Unsign...	Allow N...	Zerofill	Default	Comment
1	bc_id	INT	10	☐	☐	☐	AUTO_INCREMENT	城市id
2	bc_province	VARCHAR	10	☐	☑	☐	NULL	省
3	bc_city	VARCHAR	200	☐	☑	☐	NULL	市
4	bc_en	VARCHAR	200	☐	☑	☐	NULL	城市拼音
5	bc_first	TINYINT	4	☐	☑	☐	0	优先值

图 3 城市表

#	Name	Datatype	Length/Set	Unsign...	Allow N...	Zerofill	Default	Comment
1	bs_id	INT	10	☐	☐	☐	AUTO_INCREMENT	
2	bc_id	INT	10	☐	☑	☐	NULL	城市id
3	bs_name	VARCHAR	50	☐	☑	☐	NULL	博物馆名称
4	bs_address	VARCHAR	255	☐	☑	☐	NULL	地址
5	bs_telephone	VARCHAR	50	☐	☑	☐	NULL	电话
6	bs_uid	VARCHAR	25	☐	☑	☐	NULL	特征id
7	bs_tag	VARCHAR	50	☐	☑	☐	NULL	标签
8	bs_lat	VARCHAR	25	☐	☑	☐	NULL	经度
9	bs_lng	VARCHAR	25	☐	☑	☐	NULL	纬度
10	bs_rating	VARCHAR	2	☐	☑	☐	NULL	星级
11	bs_starttime	VARCHAR	200	☐	☑	☐	NULL	开馆时间
12	bs_price	VARCHAR	25	☐	☑	☐	NULL	价格
13	bs_time	VARCHAR	25	☐	☑	☐	NULL	时间
14	bs_season	VARCHAR	255	☐	☑	☐	NULL	季节
15	bs_photo	VARCHAR	100	☐	☑	☐	NULL	图片1
16	bs_photo1	VARCHAR	100	☐	☑	☐	NULL	图片2

图 4 博物馆表

#	Name	Datatype	Length/Set	Unsign...	Allow N...	Zerofill	Default	Comment
1	ph_id	INT	10	☐	☐	☐	AUTO_INCREMENT	
2	ph_title	VARCHAR	255	☐	☑	☐	NULL	作品名称
3	ph_author	VARCHAR	50	☐	☑	☐	NULL	作者
4	ph_ltime	INT	10	☐	☑	☐	NULL	馆藏时间
5	ph_dtime	INT	10	☐	☑	☐	NULL	入数据库时间
6	ph_material	VARCHAR	50	☐	☑	☐	NULL	材质或画风
7	ph_size	VARCHAR	50	☐	☑	☐	NULL	尺寸（以 12,12 或 12,12,12）格式...
8	ph_museum	VARCHAR	50	☐	☑	☐	NULL	储存地名称
9	ph_age	VARCHAR	50	☐	☑	☐	NULL	馆藏年代
10	ph_classes	VARCHAR	50	☐	☑	☐	NULL	艺术品类别
11	ph_location	VARCHAR	50	☐	☑	☐	NULL	储存地地区
12	ph_heritage	VARCHAR	50	☐	☑	☐	NULL	遗产形式

图 5 馆藏表

#	Name	Datatype	Length/Set	Unsign...	Allow N...	Zerofill	Default	Comment
1	id	INT	5	☑	☐	☐	AUTO_INCREMENT	id
2	class	VARCHAR	50	☐	☐	☐	No default	类库
3	name	VARCHAR	50	☐	☐	☐	No default	名称
4	entropy	VARCHAR	100	☐	☐	☐	No default	熵
5	energy	VARCHAR	100	☐	☐	☐	No default	能量
6	contrast	VARCHAR	100	☐	☐	☐	No default	对比度
7	correlation	VARCHAR	100	☐	☐	☐	No default	自相关
8	homogenity	VARCHAR	100	☐	☐	☐	No default	同质性
9	color	TEXT		☐	☐	☐	No default	颜色
10	element	TEXT		☐	☐	☐	No default	组成基元

图 6　特征库表

#	Name	Datatype	Length/Set	Unsign...	Allow N...	Zerofill	Default	Comment
1	be_id	INT	11	☐	☐	☐	AUTO_INCREMENT	错误id
2	bs_id	INT	11	☐	☑	☐	NULL	博物馆id
3	be_type	INT	1	☐	☑	☐	NULL	错误类型 1.404 2.300
4	be_count	INT	1	☐	☑	☐	0	错误次数

图 7　错误日志表

4　中华文化基因在线服务平台实现

4.1 运行环境

本系统实现是基于 PHP、C++语言作为后台，MYSQL 数据库开发的.

4.2 子系统实现

(1)检索系统

检索系统分为文本检索、基于内容的图像检索、基于语音的文本检索及基于地理位置的检索. 本数据库引擎借鉴了 Active Record(简称 AR)类, 并在 AR 类的基础上封装了常用的查询函数, 比如单条返回、多条返回、like、insert、update 等.

① 文本检索是输入文字或点击相关链接, 后台通过 get 方式或是 uri 方式获取用户需求, 对数据库的封装类进行查询并返回数据.

② 基于内容的图像检索是对上传图像进行特征提取, 然后基于数据库特征进行匹配, 返回近相似检索结果.

③ 基于语音检索主要通过使用 google voice 服务的 api. 前提是使用 webkit 内核的浏览器, 本文是 chrome 浏览器, 故首先通过 webkit 内核获得语音权限, 然后通过录音获取用户的原始数据, 接着对原始数据进行封装、编码, 将编码后的音频 POST 至 API, 最后 php 通过分析 API 返回的 JSON 数据返回结果.

④ 基于地理位置的检索主要分为两个模块: 第一个模块是基于百度地图的 api 调用获取当前地理位置及地图; 另一个模块是根据当前地理位置搜索本地数据库, 并将获取到的信息进行整合操作, 返回到用户界面. 用户可以根据自己的偏好点击进入相应的博物馆, 浏览博物馆的相关信息包括馆藏、开放时间、票价、地址、照片等.

系统界面如图 8 所示, 可以在百度地图中选择地点, 根据该点的经纬度显示该位置的原始景点风貌.

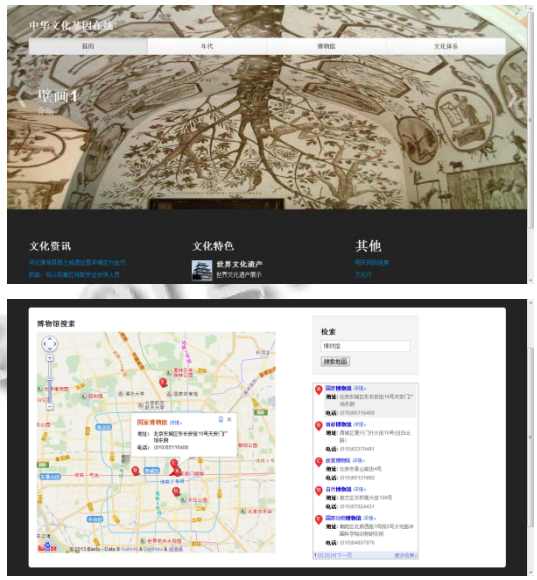

图 8　系统实现

(2)数字化系统

数字化系统是将一些本来只存在于博物馆中的文物、字画进行以照片、录影等方式进行储存, 并上传到系统中. 本文是将已经采集到好并散布于网络的图片音像文件捕捉聚合而成. 并通过算法分析, 有效提高数据分类存储.

4.3 文化基因库实现

文化基因库的实现是特征库表和博物馆表及馆藏表的实现.

(1)特征库表

通过对后台管理员批量上传相关图片进行特征提取, 构建其相应特征库表.

(2)博物馆表及馆藏表

基于不同博物馆资源完成数据化过程. 主要是数字化模型构建和基于博物馆的索引关键字.

5 小结

本文提出汇聚用户系统、检索系统、数字化系统为一体的在线服务系统. 其中检索系统包括基于文本、基于内容图像、基于语音文本以及基于地理位置的检索; 其中基于内容的图像检索只包含了纹理与颜色的特征提取及匹配; 数字化系统中的网络爬虫则使用了聚焦网络爬虫来实现. 分析文化基因在线服务系统的应用, 表明该平台对于中华文化的保护与传承具有重要价值. 平台不仅利用网络爬虫技术收集博物馆资源, 还可以起到对文化遗产数字化保护作用, 进而通过数字博物馆展示传承中华民族文化, 通过文化基因提取和建模, 实现对文化内涵的深度挖掘.

参考文献

1 邵培仁,林群.中华文化基因抽取与特征建模探索.徐州师范大学学报(哲学社会科学版),2012,2:107–111.

2 赵海英,徐正光,张彩明.一类新疆民族风格的织物图案生成方法.图学学报,2012,4:1–8.

3 赵智慧,智慧文遗——文化遗产数字化保护新理念.艺术科技, 2014,2:33–34.

4 Gonzalez-Aguilera D, Gomez-Lahoz J. Forensic Terrestrial Photogrammetry from a Single Image. Journal of Forensic Sciences, 2009, 54(6): 1376–1387.

5 Balk H, Ploeger L. IMPACT: working together to address the challenges involving mass digitizatization of historical printed text OCLC Systems & services. International Digital Library Perspectives, 2009, 25(4): 233–248.

6 路遥,王小平,苟秉宸等.基于 Web 的虚拟旅游系统.计算机辅助工程,2005,4:31–34.

7 杨安祺,殷耀文.秦郑国渠数字文化遗产再现技术.陕西科技大学学报,2006,5:112–115.

8 李向阳,庄越挺,潘云鹤.基于内容的图像检索技术与系统.计算机研究与发展,2001,3:344–354.

9 黄祥林,沈兰荪.基于内容的图像检索技术研究.电子学报, 2002,7:1065–1071.

10 Swain MJ, Ballard DH. Color Indexing. International Journal of Computer Vision, 1991, 7(1): 11–32.

11 Stricker M, Orengo M. similarity of color images. Proc. SPIE Storage and Retrieval for Image and Video Databases, 1995, 2420: 381–392 .

12 Gong Y, Chuan CH, Guo X. Image Indexing and Retrieval based on Color Histograms . Multimedia Tools and Applications, 1996, (2):133–156.

13 Pass G, Zabih R, Miller J. Comparing Images Using Color Coherence Vectors. 4th ACM Conf. on Multimedia. Boston: ACM, 1996.

14 Haralick RM. Statistical and structural approaches to texture. Proc. IEEE, 1979, 67:786–804.

15 Tamura H, et al. Texture features corresponding to visual perception. IEEE Trans. on Systems, Man and Cybernetics, 1978,8(6):460–473.

16 Jain AK, Zhong Y, Lakshmanan S. Object matching using deformable templates. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1996,18(3):267–278.