

基于层次语义的 Web 服装图像智能采集方法

耿增民^{1,2} 商书元² 邵新艳³ 周毅灵² 马 玲³

(北京服装学院数字与交互媒体北京市重点实验室 北京 100029)¹

(北京服装学院计算机信息中心 北京 100029)² (北京服装学院服装艺术与工程学院 北京 100029)³

摘 要 以大规模智能采集互联网中的服装图像为目的,研究如何利用互联网上服装图像的伴随文本与服装图像概念之间的关联,实现自动采集各语义对应的服装图像。在 HITS(Hyperlink-Induced Topic Search)算法的基础上提出一个基于层次语义的图像采集算法 SICR(Semantic-based Image Collection Robot)。该算法在层次语义库的支持下,扩充根集与去除链接工厂页面同步进行。在爬取链接网页前,进行锚文字的相似度计算和页面内容的概念分析,舍弃不符合语义的页面,只下载满足语义的服装图像。算法克服了基于文本分析或链接分析的图像自动提取算法的不足,具有较高的准确率和召回率,实验结果证明了 SICR 算法的有效性。

关键词 图像语义,图像检索,服装图像,Web 挖掘

中图法分类号 TP391.41 文献标识码 A

Hierarchical Semantic-based Web Intelligent Fashion Image Retrieval Method

GENG Zeng-min^{1,2} SHANG Shu-yuan² SHAO Xin-yan³ ZHOU Yi-ling² MA Lin³

(Beijing Lab of Digital and Interactive Media, Beijing Institute of Fashion Technology, Beijing 100029, China)¹

(Computer Information Center, Beijing Institute of Fashion Technology, Beijing 100029, China)²

(School of Fashion Art and Engineering, Beijing Institute of Fashion Technology, Beijing 100029, China)³

Abstract Aiming at the large-scale automatic collection of fashion images from Web, this paper studied how to use as-association between the accompany text and the concept of fashion images on Web pages to collect images automatically. Based on the acquisition of semantic-based Web content and the drawbacks of HITS (Hyperlink-induced Topic Search) method, a novel SICR (Semantic-based Image Collection Robot) method was proposed to collect the fashion images from Web. The proposed method (SICR), under the support of the hierarchical semantic library, removes Link Farm page in the expansion of root set, does the similarity calculation for anchor text when crawling link pages. In addition, it makes a brief conceptual analysis of the page content before downloading images. The experimental results on the large-scale dataset have demonstrated that the proposed method can overcome the deficiency of only text or link-based analysis and improve the precision rate and recall rate of fashion image retrieval, experimental results demonstrate the effectiveness of SICR.

Keywords Image semantics, Image retrieval, Fashion image, Web mining

1 引言

随着互联网和图像处理技术的发展,图像信息在人们的生活和工作中起到越来越重要的作用。互联网上存储了海量的图像数据,并且其数量仍在高速增长。例如,在线照片分享网站 www.flickr.com 仅运作 4 年,便达到每月 4 千万人的访问量,总共上传 20 亿张图像,平均每天上传几百万张^[1]。如此众多的图像,主要是通过 Google、百度等通用搜索引擎进行基于关键词的获取。当前搜索引擎的智能化程度较低,缺乏领域知识,如输入关键词“西服”后,返回的结果却有图书“西服结构设计教程”的封面图像,显然这样的结果不是需求的服装图像。而基于内容的图像检索不但计算量大,而且受

制于模式识别技术的发展,应用范围较窄^[2,3]。广大服装工作者迫切希望能够拥有更加智能化的图像采集方法。

目前主要有两种互联网图像采集方法。第一种方法是直接将现有搜索引擎返回的图像作为训练图像集,如 Torralba 等人^[4]从 WordNet 词典中提取了 75062 个名词,然后利用包括 Altavista、Ask、Flickr、Cydral、Google、Picsearch 和 Webshots 在内的搜索引擎返回各个名词(语义)对应的图像,共采集了 79302017 幅图像。第二种方法是对现有搜索引擎返回的图像进行选择或重新排序后,得到训练图像集,如 Li 等人^[5]采用增量学习方式从 Google、Yahoo、PicSearch 等搜索引擎所返回的图像中学习相应的图像分类器;Yanai^[6]通过 Google、InfoSeek 等多种搜索引擎得到指定语义对应的图像

本文受北京市教育科学“十二五”规划重点课题(AJA11174),教育部人文社科项目(12YJA760014),北京市教育委员会专项资助,北京服装学院创新团队与优秀人才选拔与培养计划项目资助。

耿增民 男,博士,副教授,主要研究领域为数据挖掘、图形图像处理,E-mail:zmgeng@gmail.com;耿增民(1968—),男,博士,副教授,主要研究方向为数据挖掘、图形图像处理,E-mail:zmgeng@gmail.com;商书元(1963—),教授,主要研究方向为虚拟现实;邵新艳(1979—),女,副教授,主要研究方向为服装工程;周毅灵(1962—),女,副教授,主要研究方向为服装计算机应用;马 玲(1980—),女,讲师,主要研究方向为服装工程。

及其网页,然后利用文本与视觉分析从中选择相关图像,最后对新图像进行 k-近邻分类。文献[7,8]对图像采集的研究也做了综合论述和介绍。总的说来,现有的方法在互联网图像采集问题上取得了一些进展,但采集效率低下,误差较高,不能较好地满足实际应用的要求,专门针对服装图像的研究更是少见。

服装图像采集是服装图像检索的重要组成部分,Web 服装图像采集的第一步是获取高质量的 Web 服装网页,只有在高质量网页中才能采集到高质量的图像。仅仅简单根据网页中图像标签的关键词来提取图像是导致当前图像挖掘准确率低下的主要原因。

互联网挖掘的两个典型算法是 PageRank^[9]和 HITS^[10],前者以整个互联网为挖掘对象,代价较高,后者以种子网站为挖掘目标,逐步扩大挖掘范围,故 HITS 更适用于面向某类主题或应用的特定领域挖掘。当前已利用 HITS 算法成功进行了舆情监测、计量信息、中医药信息等领域的挖掘^[11-14]。

互联网网页中主题相关的页面通常聚积在一起,但由于链接的跳转,容易产生主题漂移(Topic Drift)问题^[15]。分析用户在 Web 搜索时输入的信息,提炼出用户信息的主题,然后按照该主题从 Web 中发现与用户主题相关的目标网页的过程叫主题蒸馏(Topic Distillation)^[16]。但如何确定页面主题是一大研究难点,现有的方法没有充分结合领域知识,存在较大的不确定性。

若能利用 HITS 算法并结合领域知识进行主题判断,再辅以有效的页面分析如互联网上的图像的伴随文本、图像元数据及文本与图像语义之间存在的关联等,进行有选择的 Web 图像采集,有可能获得较高的服装图像采集准确率。

2 HITS 算法存在的问题

HITS 是一个面向领域的数据挖掘算法,在互联网图像搜索中有着广泛的应用。然而,该方法存在两方面的问题:1) HITS 的链接分析是局部分析,与 PageRank 进行的全局分析不同,因此一旦增加某些节点的边后,会导致页面的 Authority 和 Hub 值产生较大的变化^[17],比如 Link Farm(链接工厂)页面,本身没有多少具体内容,但却指向包括很多权威的网站,从而造成网页的 Rank 值较大,这类不合理的网页属于垃圾网页,应该从中剔除^[18];2) 若一个网站的内部网页之间人为地增加相互间链接,也会导致不应有的权值增加^[19]。

HITS 还有可能出现查询返回的结果页面和查询并非是一个主题,即主题漂移。这是因为算法在递归叠加时使得非相关页面的 Authority 和 Hub 值都比较高从而导致其出现在查询列表中^[20]。如查询“法国风格的西装”时,得到的结果却有許多旅游的主页,原因是和“法国风格的西装”有关的网页通常会链接向介绍法国的网页。

在克服 HITS 算法缺陷的基础上,增加对网页内容信息的利用,尤其是图像周边文本及网页语义的利用,提出了基于层次语义的 Web 图像的采集方法。

3 基于层次语义的网络服装图像爬取

为了提高图像检索的精度,要求搜索引擎“理解”用户的搜索意图,亦使搜索引擎具备语义知识才能采集到满足需求的图像。在构建服装层次语义知识库的基础上设计了图像爬取算法 SICR(Semantic-based Image Collection Robot)。

3.1 建立服装层次语义知识库

服装层次语义知识库可以为图像采集提供额外的主题知

识,从而提高图像自动采集的智能水平^[21]。除了包含专业词表、同义词表、领域专家名、机构名表、服装特殊符号表外,服装层次分类树图还包含了丰富的服装语义。图 1 是部分服装层次分类树图。

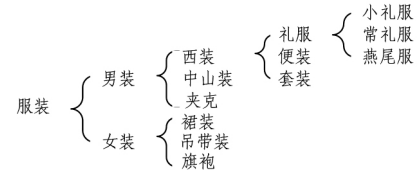


图 1 服装层次分类树图(部分)

3.2 服装语义的表示

为了表示图 1 所示的语义知识,在服装专家的参与下,确定一组带有权值的而且能够代表服装领域的关键词,并将关键词提升到语义层次。因为中文存在着普遍的歧义现象,有的词有数十个概念,同义词现象也比较普遍,需要解决多义、同义以及概念间的语义关系。借鉴知网(HowNet)^[22]、概念主题爬虫^[23]建立的层次语义知识库的数据结构表示如下。

定义某关键词 t 的语义集合为 C_t ,关键词有 n 个语义, c_i 为第 i 个语义,则 $C_t = \{tc_1, tc_2, \dots, tc_i, \dots, tc_n\}$ 。对所有的关键词做同样的定义,服装词汇中“西服”、“西装”表示同一个概念,需要对这类关键词进行处理并替换成同一个词,则得到一个二维的语义矩阵 C_T 。

图 1 中越是顶层的语义越是笼统,越是底层的服装语义越明确。把语义的个数 m 作为向量的维数,每个语义的权值 w_i 作为每一维分量的大小,根据关键词在图 1 中的层次将语义的权值进行调整,则语义主题用向量表示为 $A = (w_1, w_2, \dots, w_i, \dots, w_m)$, $i=1, 2, \dots, m$ 。

3.3 页面的语义提取

对获取的 Web 页进行分析,在语义集合上定义页面 p 的特征向量为 $V_c(p) = \{c_1, w_{c_1(p)}, \dots, c_i, w_{c_i(p)}, \dots, c_m, w_{c_m(p)}\}$,其中 $w_{c_i(p)}$ 为语义 c_i 在页面 p 中的权重,计算公式为:

$$w_{c_i(p)} = \sum_j w_{t_j c_i(p)} \quad (1)$$

其中, $t_j c_i$ 表示页面 p 中关键词 t_j 包含 c_i 语义的概率值,经过归一化处理后,计算公式为:

$$w_{t_j c_i(p)} = \frac{f_{t_j c_i(p)} \log \left| \frac{N}{n_{t_j c_i(p)}} \right|}{\sqrt{\sum_{i=1}^n f_{t_j c_i(p)}^2 \log^2 \left| \frac{N}{n_{t_j c_i(p)}} \right|}} \quad (2)$$

其中, N 为文档总数, $f_{t_j c_i(p)}$ 为页面中出现关键词的次数, $n_{t_j c_i(p)}$ 为出现关键词相关语义的文档频率。

计算出每个关键词的语义值后,对所有关键词的语义值进行归一化合并,得到页面语义向量 $p = (v_1, v_2, \dots, v_i, \dots, v_m)$ 。

3.4 查询语义和页面语义的相似度计算公式

用户根据需要在搜索引擎中输入查询句,再将查询句进行关键词抽取,则查询语义向量定义为 $R = (r_1, r_2, \dots, r_i, \dots, r_m)$,其中 r_i 为第 i 个语义的权重, R 和 P 的语义相关度计算公式为:

$$Sim(P, R) = \frac{\sum_{i=1}^m v_i r_i}{\sqrt{\sum_{i=1}^m v_i^2} \sqrt{\sum_{i=1}^m r_i^2}} \quad (3)$$

指定一个阈值,当 $Sim(P, R) \geq$ 阈值时就认为该页面和查询是相关的,阈值的取值需要根据经验和实际测试确定,若

想获取较多的页面,可以把阈值设小一点,反之则可以把阈值设大一点。

3.5 服装图像的爬取

只有当网页和查询主题相关时,网页中的图像才可能是用户需求的图像。通常利用 HTML 网页中的 `<img>` 标签、`<alt>` 标签对图像进行定位、下载。值得注意的是对于背景图像需要查询关键字“background”或查询图像文件的名字后缀“.gif”、“.jpg”、“.bmp”等。

即使网页的主题是与服装相关的,也不能保证网页中的所有图像都是用户感兴趣的,如某个服装公司的网页里面也许有公司厂房的图像,这时就需要程序通过大量的机器学习进行甄别,只下载用户有可能感兴趣的图像。

3.6 算法描述

本文提出的 SICR 算法试图根据用户输入的查询理解并确定用户的兴趣领域。该算法的流程图如图 2 所示。需要说明的是:

(1)首先精心选择一些服装类网站(如大的品牌公司、时装秀网站、流行趋势网站)作为种子网站,然后在爬取网页的过程中对 HITS 算法进行了一定的改进,以确保采集到的网页是与用户的查询主题相关的,并且网页中存在服装图像。

(2)算法首先用多智能主体的 Spider 以基于语义的形式爬取得到根集(Root Set)。由于采用了优于关键词的概念页面主题提取方法,使得获取的根集要优于基于关键词获取的根集。

(3)在扩充根集 R 的过程中,HITS 直接将页面包含的 URL 加入到基本集 S 中,这有可能将一些与主题无关的页面也扩充进来了。算法首先剔除了 Link Farm 页面;然后对 URL 的 Anchor Text 进行了分析,若其与主题相关,直接将其扩充到 S ,否则再提取页面概念并进行相关度计算。虽然以时间付出为代价,但是换来了整个提取页面的质量的提高。假设页面长度为 n ,对页面文本的处理仅限于关键词的提取,

在关键词集合不大的时候,算法的时间复杂度接近于 $O(n)$ 。

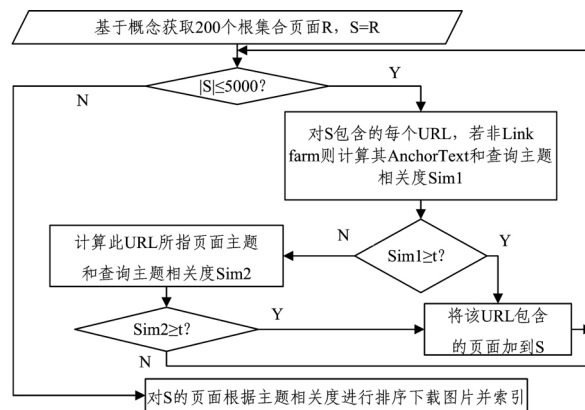


图 2 SICR 算法流程图

4 实验及分析

实践中,对服装图像的采集方法一般是基于网页中图像的标签采集或利用 Google 等搜索引擎返回图像^[24,25],这样采集的图像召回率和准确率会受到一定的影响。

本文的实验主题是基于纺织服装领域的,搜索深度为 2,起始种子为 20(与纺织服装紧密相关的国内外权威网站个数),阈值设为 0.15。查询关键词包括 30 个中文关键字和 10 个英文关键字,实验结果只列出了 6 个关键字,分别是“燕尾服”、“连衣裙”、“真丝”、“开衫”、“Cardigan”、“Tuxedo”。

考虑到 Google 搜索引擎被公认为目前查询精度最高的搜索引擎,所以把实验结果和 Google、HITS 算法进行了比较(Google 数据于 2015 年 10 月 22 日获取)。由于 SICR 算法目前仅索引 com 和 org 网址,为了保证对比的公正性,在 Google 检索中使用了域名限制为“site:com”和“site:org”,对所有图像进行了人工评判。

实验数据由表 1 所列。

表 1 SICR 和 Google、HITS 的实验比较

类别	获取方式	燕尾服	连衣裙	真丝	开衫	Cardigan	Tuxedo	平均正确率
Google	基于关键词	1422	1479	1489	1456	1301	1067	83%
HITS	基于关键词	236	383	175	391	217	294	89%
SICR	基于概念	375	391	260	478	433	453	96%

由实验数据可以看出,由于 HITS 和 SICR 算法是基于主题的搜索,再加上硬件规模的限制,返回图像的数量不如 Google 多。但 HITS 和 SICR 召回的图像内容属于纺织服装领域的准确度更高。由于有层次语义库的支持,SICR 召回的图像数量较 HITS 稍多一些,这是因为在检索“西装”图像时,与“西装”相关的概念如“三件套”、“制板”、“纸样”等图像也被召回。在获得较高准确率的基础上,扩大程序的规模,专门基于社交网络媒体的图像进行挖掘和对流行趋势的预测,正是我们下一步要做的工作。

参考文献

- [1] Paulus M J, Ingersoll R. Digital Wisdom for a Digital Age: Spirituality and Technology in the 21st Century[C] // Christ and Cascadia, 2014
- [2] Babenko A, Slesarev A, Chigorin A, et al. Neural codes for image retrieval[M] // Computer Vision-ECCV 2014. Springer International Publishing, 2014: 584-599
- [3] Wang R X, Peng G H, Zheng H C. A New Image Retrieval Al-

gorithm Based on Sparse Coding[C] // 2015 International Conference on Artificial Intelligence and Industrial Engineering. Atlantis Press, 2015

- [4] Pandey M, Lazebnik S. Scene recognition and weakly supervised object localization with deformable part-based models[C] // 2011 IEEE International Conference on Computer Vision (ICCV). IEEE, 2011: 1307-1314
- [5] Li L, Li F. OPTIMOL: Automatic online picture collection via incremental model learning[J]. International Journal of Computer Vision, 2010, 88(2): 147-168
- [6] Cheng Qi-min, et al. Combining SIFT and global features for Web image classification[M] // Advances in Multimedia Information Processing-PCM 2012. Springer Berlin Heidelberg, 2012: 739-747
- [7] Pereira R, Lopes L S, Silva A. Semantic image search and subset selection for classifier training in object recognition[M]. Progress in Artificial Intelligence. Springer Berlin Heidelberg, 2009: 338-349
- [8] 张磊. 大规模互联网图像检索与模式挖掘[J]. 中国科学信息科

学(中文版),2013,43(12):1641-1653

- [9] Ishii H, Tempo R, Bai E W. A web aggregation approach for distributed randomized PageRank algorithms[J]. IEEE Transactions on Automatic Control, 2012, 57(11): 2703-2717
- [10] Singh A, Sharma S. Role of Page ranking algorithm in Searching the Web; A Survey[J]. International Journal of Engineering & Technology, Management and Applied Sciences, 2014, 1(1): 39-43
- [11] Lappas T, Liu Kun, Terzi E. A Survey of Algorithms and Systems for Expert Location in Social Networks[M]// Social Network Data Analytics. Springer US, 2011: 215-241
- [12] 吴少华, 崔鑫, 胡勇. 基于 SNA 的网络舆情演变分析方法[J]. 四川大学学报(工程科学版), 2015, 47(1): 138-142
- [13] Dewgun T K, Chauhan P S. A Survey on Web usage Mining: Process, Techniques and Applications[C]// International Journal of Engineering Research and Technology. ESRSA Publications, 2015
- [14] 孙颖伟, 睦蕴慧, 张磊, 等. 数据挖掘技术在中医辨证规律研究中的应用进展[J]. 北京中医药, 2015, 34(1): 70-74
- [15] Liu Xin-yue, Lin Hong-fei, Zhang Cong. An Improved HITS Algorithm Based on Page-query Similarity and Page Popularity [J]. Journal of Computers, 2012, 7(1): 130-134
- [16] Wu Ming-fang, Hawking D, Turpin A, et al. Using Anchor Text for Homepage and Topic Distillation Search Tasks[J]. Journal of the American Society for Information Science and Technology,

2012, 63(6): 1235-1255

- [17] Alsoos M, Kheirbek A. A Semantic Approach to Enhance HITS Algorithm for Extracting Associated Concepts using ConceptNet [J]. Journal of Digital Information Management, 2015, 13(1): 55
- [18] Amin M S, Kabir S, Kabir R. A Score based Web Page Ranking Algorithm[J]. International Journal of Computer Applications, 2015, 110(12): 11-15
- [19] Saraswathi D, Kavitha R. A new enhanced technique for link farm detection[C]// 2012 International Conference on Pattern Recognition, Informatics and Medical Engineering (PRIME). IEEE, 2012: 74-81
- [20] Hatakenaka S, Miura T. Query and topic sensitive pagerank for general documents[C]// 2012 14th IEEE International Symposium on Web Systems Evolution (WSE). IEEE, 2012: 97-101
- [21] Geng Zeng-min, Li Xue-fei, Liu Yu-yu. Design of an Intelligent Fashion Information System[C]// TBIS 2011 Advanced Textiles, Fashionable industry. Beijing, China, 2011: 27-29
- [22] 董振东, 董强. 知网[EB/OL]. <http://www.keenage.com>, 2013
- [23] 黄仁, 王良伟. 基于主题相关概念和网页分块的主题爬虫研究[J]. 计算机应用研究, 2013, 30(8): 2377-2380
- [24] 万玉钊, 刘峡壁, 韩菲罪, 等. 用于提高谷歌图像搜索结果的二分类器在线学习方法 [J]. 自动化学报, 2014, 40(8): 1699-1708
- [25] Zhou N, Fan J. Automatic image-text alignment for large-scale Web image indexing and retrieval [J]. Pattern Recognition, 2015, 48(1): 205-219

(上接第 241 页)

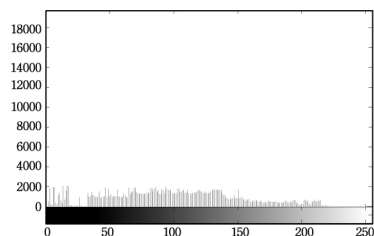


图 5 压缩后图像的灰度图

图 4 为原始图像的灰度图,图 5 为图像压缩后的灰度图。可以看出,与图 4 相比,图 5 具有较少的信息,可以得到更均匀的图像,压缩后图像的效果比较理想。

结束语 本文主要研究了利用 DCT 技术对图像进行压缩,通过 Matlab 仿真分析了利用 DCT 技术进行图像的压缩的优点。随着网络媒体的发展,图像压缩技术将进入快速发展时代,尤其是图像压缩标准的建立使它进入实用化和产业化。使用 DCT 进行图像压缩具有更高的性能,并且它易于实现。DCT 具有算法简单、易于硬件实现等优点,从而广泛应用于图像压缩领域,尤其在较高比特率的传输条件下能取得令人满意的压缩效果。然而,它在低比特率条件下的“方块效应”成了不容忽视的“瓶颈”缺陷,其原因在于 DCT 处理图像时是分块进行的,块与块边界两边的像素相关性不能去掉,从而形成了明显的方块边界。若将 DCT 变换与其它编码方法(如小波变换、分形等)相结合将得到更好的效果。

参考文献

- [1] 张爱华,江中勤,张华. 基于粒子群优化算法的分形图像压缩编码[J]. 计算机技术与发展, 2010, 20(2): 21-24

- [2] Narwaria M, Lin W S. Objective image quality assessment based on support vector regression[J]. IEEE Transactions on Neural Networks, 2010, 21(3): 515-519
- [3] 孙秉钰. Huffman 算法的改进及其在图像压缩中的应用[J]. 微计算机信息, 2011, 27(1): 264-266
- [4] 武善玉, 晏振鸣. 改进的 Huffman 编码及其应用[J]. 通信技术, 2009, 42(1): 309-311
- [5] 朱艳秋, 初练禹, 陈贺新. 一种基于二维 DCT 的分形静止彩色图像压缩编码[J]. 中国图象图形学报, 1997, 2(7): 486-487
- [6] 罗晨. 基于 DCT 的图像压缩及 MATLAB 实现 [J]. 电子设计工程, 2011, 19(18): 220-224
- [7] 张兆礼, 赵春晖, 梅晓丹. 现代图像处理技术及 MATLAB 的 DCT 变换在 JPEG 图像压缩中的应用[M]. 北京: 人民邮电出版社, 2001
- [8] 陈波, 王红霞, 成礼智. 图像压缩中的快速方向离散余弦变换 [J]. 软件学报, 2011, 22(4): 826-832
- [9] 张智高, 春花, 张红梅, 等. 离散余弦变换在图像压缩上的应用 [J]. 内蒙古民族大学学报(自然科学版), 2010, 25(3): 153-155
- [10] 高国芳. 数字图像压缩处理技术[J]. 重庆科技学院学报(自然科学版), 2006, 8(2): 93-95
- [11] 张健. 基于离散余弦变换数据压缩算法的图像处理应用[J]. 科技咨询导报, 2007, 8(42): 57-58
- [12] 方建斌. 变换编码在图像压缩中的应用[J]. 中国水运: 下半月, 2008, 8(12): 109-110
- [13] 吕世良, 曲仕敬. 基于 DCT 的图像压缩及 Matlab 实现 [J]. 科技信息: 科学教研, 2008, 2(14): 401-402
- [14] 王光伟. 一种离散余弦变换压缩算法的改进[J]. 计算机与信息技术, 2008, 1(8): 15-18
- [15] 张浩, 杨永健, 韩鸿莺. 一种基于 DCT 变换的序列分组水印法 [J]. 计算机应用研究, 2014, 20(2): 46-48