EMS表单中手写体中文识别图像预处理方法研究

许秦蓉
（上海理工大学，上海 200093）

摘要：目的 在脱机手写体文字识别系统中，由于自由书写的字符不可避免地受到图像背景不均匀、图像倾斜和字符粘连及大小不一等因素的影响，为了确保字符切分和识别的正确性，对EMS表单中手写体汉字字符图像预处理方法进行探讨，展示了EMS表单图像预处理的全过程。**方法** 采用最小二乘法作拟合直线的方法，对目标图像进行定位和分割，用基于大津阈值的分块阈值算法处理目标图像的背景不均问题，并减少噪声干扰。**结果** 该图像预处理方法在1020张真实EMS图像上进行测试，识别正确率达到了86.3%。**结论** 该方法有一定的灵活性和抗干扰性，减少了图像噪声对汉字字符切分和识别的影响。

关键词： 手写中文字符；识别；图像分割；图像预处理
中图分类号： TP391.41 **文献标识码：** A **文章编号：** 1001-3563(2014)21-0080-06

Image Preprocessing Method for Recognition of Handwritten Chinese Characters in EMS Forms

XU Qin-rong
(University of Shanghai for Science and Technology, Shanghai 20093, China)

ABSTRACT: Objective In OCR system, image preprocessing is particularly important for recognition of unlimited handwritten Chinese characters. Some unavoidable factors from image background, image skew and touching characters bring in errors for character segmentation, recognition and post-processing. In this paper, we focused on the image preprocessing method for recognition of handwritten Chinese characters in EMS Forms and the whole process was shown. **Methods** Method of finding fitting Straight Line by Least Squares was used to deal with the relocation and segmentation of the target image. Block threshold based Otsu´s method was adopted to remove the image background and to eliminate the noise interferences. **Results** The proposed method was tested on 1024 real EMS envelope images. The system achieved a recognition rate of 86.3%. **Conclusion** The experimental results demonstrated that the proposed method effectively reduced the influence of image noises on the segmentation and recognition of handwritten Chinese characters.

KEY WORDS: handwritten Chinese characters; recognition; image relocation and segmentation; image preprocessing

脱机手写体文字识别是指通常所说的光学文字识别(Optical Character Recognition，OCR)。目前，印刷体和联机手写体汉字识别已经实用化，但脱机手写体汉字识别仍停留在研究阶段，究其原因，主要是识别效果很大程度依赖于对目标汉字图像的预处理效果，包括图像的去除背景、倾斜校正、消除噪声和汉字及数字的切分等。尽管脱机手写体汉字的识别是汉字识别中难度最大的，但如果预处理效果好，能够正确切分字符(包括区分汉字和数字)，使用汉字识别模块基本上都能得到较正确的候选字，之后的难点集中在汉字的后处理上，即如何从候选字集中找出正确的候选字。然而，对于脱机手写体汉字图像的预处理和字

收稿日期： 2013-10-07
作者简介： 许秦蓉(1973—)，女，陕西人，在读博士，上海理工大学讲师，主要研究方向为印刷包装工程及模式识别。

符切分,特别是非限定人、自由书写的汉字和数字,不可避免地会出现图像背景不均匀、图像倾斜和字符粘连及大小不一等情况,这些干扰信息都会造成图像预处理和字符的切分错误,错误切分的字符会直接导致字符识别和后处理错误[1—3]。

在此,探讨EMS表单中手写体汉字的图像预处理方法,给出了整个中文手写体识别系统的概述和流程图,展示了EMS表单图像预处理的过程,包括目标图像定位、目标图像的背景处理、噪声处理和倾斜校正。

1 系统构架

建立地址库L,其中的所有地址都是由中国邮政提供的有效邮政地址。采用基于地址的驱动方法对手写中文邮政地址进行识别和后处理。整个过程的系统图见图1。

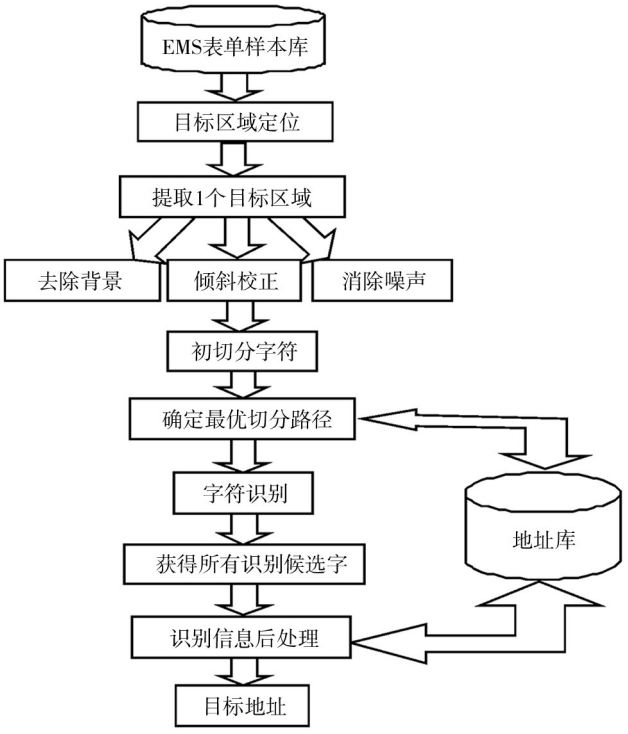

图1 系统示意

Fig.1 The system schematic diagram

EMS表单是标准化的邮政信函,其中的字符信息在提取前需要对目标进行定位,提取出目标区域图像,并对目标图像进行预处理,包括图像去除背景(二值化)、去除噪声和倾斜校正等。系统使用的是单个汉字识别系统,需要对预处理后的目标图像进行单个汉字的分割后才能识别。字符识别过程被分为粗选和精选2个阶段,在粗选阶段,所有初始切分字符及其组合节点使用汉王OCR系统识别,可获得相应的识别候选字,所有的识别候选字组成1个Trie结构。对应于地址库L中的每1条地址,在Trie结构中寻找出现的字符,当出现字符数与该条地址字符数的比率超过预先设定的阈值时,该条地址作为候选地址被放入新地址库L′。粗选结果一方面可以把地址库L中与图像地址相似度较低的地址过滤掉;另一方面把Trie结构中对应于出现字符的合理节点也选了出来,并且可以把其他不需要的节点和相应识别字符去除;在精选阶段,采用非精确字符匹配技术来评估地址库L'中的候选地址,具有最高相似度的地址就是目标地址,也即是图像地址的识别结果。

2 EMS表单图像预处理

对EMS表单中的字符信息进行提取前,需要先对图像进行预处理,包括目标定位和分割、图像二值化、图像去除噪声和图像倾斜校正等4个步骤。

2.1 目标区域定位与分割

目标定位是指EMS表单中各个目标区域的位置。EMS表单虽然格式比较固定,每个需要提取的目标区域都有明确的边框,但由于表单本身比较小,人们在填写时经常有文字信息出界,或字符与边框重叠的现象发生,这些干扰对字符正确切分及信息提取都有较大的影响。真实的EMS表单图像见图2。

图2 EMS表单图像

Fig.2 An EMS form image

2.1.1 EMS表单图像定位

该系统中处理的EMS表单是从邮政分拣机上拍摄的照片,图像中的黑色背景是分拣机的传输皮带。由于拍摄的EMS表单图像外带有黑色边框,为了去掉

此黑色边框，利用边框与EMS表单图像颜色值差异较大的特性，先对图像做简单的二值化处理。二值化后，EMS表单图像内部基本成为白色，寻找黑色边框和EMS表单图像的黑白交界点，对这些交界点作拟合直线。在EMS表单图像的上下左右作4条拟合直线，即为EMS表单图像的边界。4条拟合直线的两两交点即为EMS表单图像的4个顶点。

采用最小二乘法求边界交点的拟合直线。EMS表单图像的尺寸为2048像素×1536像素，在其边界均匀取50~ 100个点，可以得到一系列成对的数据(x_1,y_1;x_2,y_2;...;x_{50},y_{50})，将这些数据描绘在xOy平面中，可以粗略看出这些点大致散落在某直线近旁，因此认为x与y之间近似满足线性函数关系。假设直线为$y=ax+b$，寻求这50个定点到该直线距离的平方和为最小，即最小二乘解，由此求得该直线的斜率a和截距b：

$$\begin{cases} a = \dfrac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \\ b = \dfrac{1}{n}\sum_{i=1}^{n} y_i - \dfrac{a}{n}\sum_{i=1}^{n} x_i \end{cases} \quad (1)$$

EMS区域定位见图3。其中，最外侧的大矩形框即为由4条拟合直线确定的EMS表单图像边框。图像边界定位也可以采用直接获取直线特征的方法，但在图像采集过程中，有时会出现图像印刷的位置偏移或图像破损等情况，造成无法提取直线特征来确定边界。采用最小二乘法求边界交点拟合直线的方法，选取离散随机的边界点可以获得较好的定位效果，见图4。

图3 EMS区域定位

Fig.3 Relocation of EMS image

图4 破损EMS图像定位

Fig.4 Relocation of damaged EMS image

2.1.2 EMS表单各目标区域定位

由于EMS表单格式固定，当确定了其图像的边界后，就可以根据相对位置确定各目标区域的位置。考虑到EMS图像本身的问题（如图像倾斜、破损、印刷的位置偏移等），如果仅依赖图像的4条边界来定位表单中的各个目标区域，定位不够精确。因此，在EMS图像的指定区域内寻找图像中的固有信息，对照和验证图像位置。在图像的“国内特快专递邮件详情单”字样中，采用寻找连通元的方法，找到“单”和“递”2个字，以此进一步确定图像的相对位置，根据该相对位置，按照比例大小得到各目标区域较准确的位置。

EMS表单中包含15个目标区域，如条码及收寄件人的姓名、单位地址、地址、电话、城市和邮编等。 在图3中，小矩形框分别是EMS表单各目标区域的位置。分割出的收件人地址图像见图5。

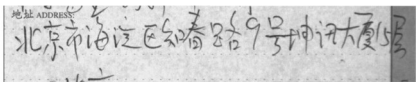

图5 分割出的地址图像

Fig.5 The block image of postal address

2.2 目标图像背景处理

目标区域图像分割出来后，需要对图像进行去除背景处理，即设定某一阈值将图像的像素分成2部分，一部分表示前景的像素，即目标区域；另一部分代表背景的像素，即背景区域，这种前景和背景的分割可以看作是图像二值化处理的一种。图像必需经过二值化处理后，才能提取前景图像中的手写字符信息。

选取阈值是一种区域分割技术，当图像前景与背景对比强烈，通过观察图像直方图可以简单地设定阈值。当图像背景比较复杂时，阈值选择却是很困难的问题。因为在数字化的图像数据中，无用的背景数据和目标对象的数据常常重叠在一起。在实际应用中，如在车牌识别或OCR文字识别中，由于光照不均匀、图像噪声等原因，会造成图像的二值化效果不理想，导致部分文字丢失或者被黑色区域掩盖，从而造成后续的字符识别结果不理想。同时，图像本身的背景颜色不均匀，或不同区域采用的不同颜色背景（如EMS表单图像），也使得图像进行二值化处理时需要区别对待，不能选用单一阈值划分目标区域和背景区域。

常用的阈值法分为全局阈值法和局部阈值法[5—6]。全局阈值法是指在二值化过程中整幅图像只使用1个全局阈值T的方法。典型的全局阈值方法包括平均灰度值算法、大津阈值算法、迭代阈值算法和最小错误

阈值算法等。很明显,对于该系统中的EMS表单图像使用全局阈值法是不合适的,应采用局部阈值法来处理此类图像;局部阈值法是用像素灰度值和此像素邻域的局部灰度特性来确定该像素的阈值。当光照不均匀、有突发噪声,或者背景灰度变化较大时,局部阈值法根据像素的坐标位置关系自动确定不同阈值,实施动态的自适应二值化处理。比较典型的局部二值化算法有Bernsen算法、Niblacks算法和基于大津阈值的分块阈值算法等。这里,采用基于大津法的分块阈值算法处理EMS表单图像。

大津法(简称OTSU)又称为最大类间方差法,它是按图像的灰度特性将图像分成背景和目标2部分。背景和目标之间的类间方差越大,说明构成图像2部分的差别越大,当部分目标错分为背景或部分背景错分为目标都会导致2部分差别变小。对于图像$I(x,y)$,前景(即目标)和背景的分割阈值记作T_t,属于前景的像素点数占整幅图像的比例记为n_1,其平均灰度为m_1;背景像素点数占整幅图像的比例为n_2,其平均灰度为m_2,类间方差记为g。假设图像的大小为$M\times N$,根据公式(2)遍历目标图像计算方差,取方差最大$\max\{g,g\in(0,255)\}$时的灰度值为阈值T_t。

$$g=n_1\times n_2\times(m_1-m_2)^2 \tag{2}$$

基于大津阈值的分块阈值算法是先将图像分成若干个大小的子图像,在每个字图像中采用大津阈值法。图像分块大小的选择将对图像的二值化结果有很大影响。如果划分的小块太小,则有的小块全部是目标或是背景,这样计算出的局部阈值将是一种强制性的结果而不是合理的阈值选择;如果划分太大,则二值化的效果又不理想。因此,划分的原则是尽量使每小块内既包含目标像素又包括背景像素,而且目标与背景各自内部的灰度等级范围较小。由实验得知,在EMS表单图像中选择5像素×5像素或10像素×10像素大小的窗口,可以获得较好的分割效果。采用基于大津阈值的分块阈值算法避免了单一阈值对复杂背景图像二值化效果的影响。4种算法的二值化图像见图6。其中,图6a是从EMS表单图像中分割出的收件人姓名图像,图6b—e是按照上述4种算法处理的二值化效果。由实验结果可知,采用基于大津阈值的分块阈值算法处理效果比较好。

除了图像中的背景颜色外,EMS表单中还有一些不需要的印刷提示信息,如“收件人姓名TO”、“单位名称COMPANY NAME”等。按照前面划定的目标区域,见图2中的小矩形框,把这些文字直接作为背景处理掉了。但该方法对于手写文字和提示信息重叠的情况,会导致字符笔画丢失。目标图像的预处理见图7。图7a是从EMS表单图像中分割出的收件人姓名图像,图7b是采用上述方法去除背景后的图像效果。

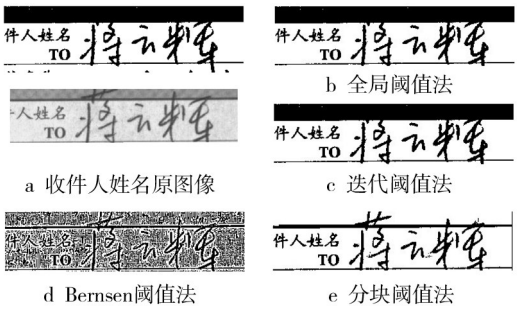

图6　4种算法的二值化图像

Fig.6 Binarization image based on four algorithms

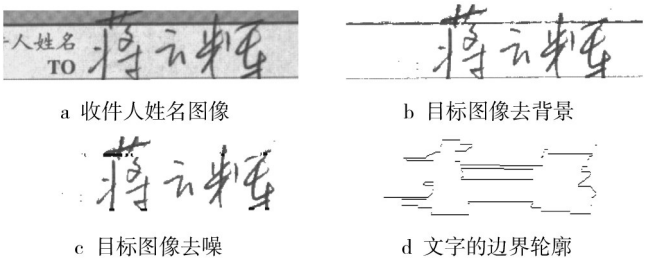

图7　目标图像的预处理

Fig.7 Preprocessing of target image

2.3　目标图像噪声处理

数字图像的噪声除了一些突发噪声外,图像本身也包含对提取字符文字信息有干扰的噪声,如书写线、边框线等。EMS表单的每个区域都有固定的书写线和边框,书写的字迹常常有与边框线交叉重叠的情况。为了保证提取出的字符信息正确,在去除图像背景后,还需要专门去除书写线和边框线。

鉴于EMS表单是固定格式,书写线和边框的位置在目标区域定位时就可以确定。书写线和边框线都是直线,确定书写线的上下边界后,即可把书写线部分做背景处理。但如果书写线与汉字笔画有重叠,直接去除书写线会造成字迹笔画断裂而影响后续的文字识别,因此,还需根据笔画的走势还原部分书写线,保证汉字笔画的连贯性。首先设定书写线宽度阈值T_w,宽

度阈值设定可以采用2种方法设定：其一，书写线是图像中的固有信息，可根据图像中书写线的粗细自行设置，例如，在EMS表单中设定为6个像素，在书写线的上下各扫描3个像素；其二，作为文字的位置边界，从视觉上来看书写线的宽度和手写体文字的笔画宽度近似，因此可参考文字的笔画宽度设定书写线宽度阈值。在文字信息区中以2×2窗口扫描图像，要求在每个窗口中有至少2个黑色相邻像素，记录黑色像素和白色像素的个数，根据公式(3)求得文字笔画宽度。

文字笔画宽度=黑色像素个数/(黑色像素个数-白色像素个数) (3)

以书写线位置为中心，沿垂直方向在书写线上下搜寻连通元。当连通元的数目小于宽度阈值T_w时，就认为是书写线，作为背景处理；否则认为属于汉字笔画，予以保留。图7c是去除书写线和边框后的文字效果，可以看出“蒋”字草头下方的横线是被保留的部分书写线，这部分保留的书写线不会影响该字的识别结果。

2.4 目标图像倾斜校正

由于该系统主要处理的是手写体汉字，因此在做倾斜校正时，并不对目标区域图像本身，而是直接对图像中的手写体汉字做倾斜校正，这样对后续的汉字切分有很大帮助，并且可以简化倾斜校正时的工作量。对1行汉字的倾斜校正方法，是先对图像做膨胀处理，寻找该行汉字的边界轮廓，即寻找前景和背景的交界点，对这些交界点作拟合直线，并以最长1条拟合直线的角度作为该行汉字的倾斜角度。当其倾斜角度大于5° 时，则旋转图像进行倾斜校正。图6d是目标图像图6a中文字的边界效果，由于其倾斜角度不大于5° ，因此不需要倾斜校正。

3 结语

以EMS表单作为研究对象，在Visual C++6.0软件平台上实现对EMS表单图像的处理工作，自动完成目标区域定位和分割、图像预处理，包括图像二值化、去除噪声和倾斜校正等过程，为后续汉字字符切分、识别和识别后处理工作做好准备。由于目标图像的预处理效果会直接影响字符切分、识别和后处理的结果，如果处理不好，会导致切分的字符识别和后处理错误，因而显得尤为重要。

采用的预处理方法在1020张真实EMS图像上做了测试，为了减少由于误切分和误识别的影响，采用地址驱动方法来预测地址字符串，并用非精确字符匹配技术在地址库中搜索匹配的地址。实验结果显示了该方法有一定的灵活性和抗干扰性，识别正确率达到了86.3%。

参考文献：

[1] 李元祥. 利用上下文信息的汉字识别理论和方法的研究[D]. 北京：清华大学，2001.
LI Yuan-xiang. The Research on Chinese Character Recognition Using Contextual Information[D]. Beijing: Tsinghua University, 2001.

[2] 苏统华. 脱机中文手写识别——从孤立汉字到真实文本[D].哈尔滨：哈尔滨工业大学，2008.
SU Tong-hua. Off-line Recognition of Chinese Handwriting: From Isolated Character of Realistic Text[D]. Haerbin: Harbin Institute of Technology, 2008.

[3] 龙翀，庄丽，朱小燕，等. 手写中文地址识别后处理方法的研究[J]. 中文信息学报，2006，20(6)：69—72.
LONG Chong, ZHUANG Li, ZHU Xiao-yan, et al. A Post-processing Approach for Handwritten Chinese Address Recognition[J]. Journal of Chinese Information Processing, 2006, 20(6): 69—72.

[4] 孙洁娣，温江涛，李书茉，等. 局部高亮干扰文本图像的二值化方法研究[J]. 光电工程，2012，39(11)：75—78.
SUN Jie-di, WEN Jiang-tao, LI Shu-mo, et al. Binarization Method for Document Images with Local Highlight Interference[J]. Opto-Electronic Engineering, 2012, 39(11): 75—78.

[5] 谢国庆，白莹，王智文. 基于全局迭代阈值和局部分析的护照图像的二值化算法[J]. 计算机应用与软件，2009，26(11)：118—119.
XIE Guo-qing, BAI Ying, WANG Zhi-wen. Binarization Method of Passport Image Based on Global Iterative Threshold and Local Analysis[J]. Computer Applications and Software, 2009, 26(11): 118—119.

[6] 吴炜，骆剑承，陈亮，等. 复杂背景下粉笔数字的字符自动提取方法研究[J]. 计算机应用研究，2009，26(10)：3963—3965.
WU Wei, LUO Jian-cheng, CHEN Liang, et al. Automatic Extraction Method of Chalk Characters in Context of Complex[J]. Application Research of Computers, 2009, 26(10): 3963—3965

[7] 申静. 基于人眼视觉特性的包装印刷图像水印技术研究[J]. 包装工程,2012,33(1):113—118.
SHEN Jing. Packaging and Printing Images Watermark Technology Based on Human Vision Characteristics[J]. Packaging Engineering,2012,33(1):113—118.

[8] 徐宏平,万晓霞,许法强. 基于视觉模型和误差扩散的半色调水印算法[J]. 包装工程,2007,28(12):77—79.
XU Hong-ping, WAN Xioa-xia, XU Fa-qiang. Watermarking Algorithm for Halftone Images Based on HVS AND Error Diffusion[J]. Packaging Engineering,2007,28(12):77—79.

[9] 李孟涛,孙刘杰,张雷洪,等. 基于小波变换的傅里叶加密印刷水印算法研究[J]. 包装工程,2012,33(1):108—112.
LI Meng-tao, SUN Liu-jie, ZHANG Lei-hong, et al. Research on Fourier Encryption Printing Watermarking Algorithm Based on Wavelet Transform[J]. Packaging Engineering, 2012,33(1):108—112.

[10] 黄军. 基于离散余弦变换域的数字水印研究[J]. 包装工程,2012,31(13):108—110.
HUANG Jun. Research of Digital Watermark Based on DCT Domain[J]. Packaging Engineering,2012,31(13):108—110.

[11] 王灿才. 基于空间域LSB数字水印的鲁棒性研究[J]. 包装工程,2009,30(3):76—78.
WANG Can-cai. Study of the Robustness of Digital Watermark Based on Least Significant Bit[J]. Packaging Engineering,2009,30(3):76—78.

[12] LI Y X, TAN C L, DING X, et al. Contextual Post-processing Based on the Confusion Matrix in Offline Handwritten Chinese Script Recognition[J]. Pattern Recognition, 2004, 37(9):1901—1912.

[13] LONG C, ZHU X, HUANG K, et al. An Efficient Post-processing Approach for Off-line Handwritten Chinese Address Recognition[C]// 2006 8th International Conference on Signal Processing,2006:16—20.

[14] NAVARRO G, RAFFINOT M. Flexible Pattern Matching in Strings: Practical on-line Search Algorithms for Texts and Biological Sequences[M]. New York: Cambridge University Press USA,2002.

[15] PENG F, SCHUURMANS D. Self-supervised Chinese Word Segmentation[C]// IDA′01: Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis, 2001:238—247.

(上接第79页)

2009,58(2):952—958.

[9] 黄素娟,王杜瑶,刘晓静,等. 基于小波变换的硬拷贝全息水印[J]. 光电子·激光,2011,22(9):1415—1420.
HUANG Su-juan, WANG Du-yao, LIU Xiao-jing, et al. Hardcopy Hologram Watermarking Based on Discrete Wavelet Transform[J]. Journal of Optoelectronics · laser, 2011, 22(9):1415—1420.

[10] WANG S, HUANG S, ZHANG X, et al. Hologram-based Watermarking Capable of Surviving Print-scan Process[J]. Applied Optics,2010,49(7): 1170—1178.

[11] 王子煜,孙刘杰,李孟涛. 强鲁棒性QR码水印技术[J]. 包装工程,2012,33(15):84—87.
WANG Zi-yu, SUN Liu-jie, LI Meng-tao. QR Code Watermark Technology with Strong Robustness[J]. Packaging Engineering,2012,33(15):84—87.

[12] 孙刘杰,李孟涛. 基于CMYK颜色空间的光全息水印算法研究[J]. 包装工程,2012,33(9):27—32.
SUN Liu-jie, LI Meng-tao. Study on Light Holographic Watermarking Algorithm Based on CMYK Color Space[J]. Packaging Engineering,2012,33(9):27—32.

[13] 周中原,孙刘杰,唐波,等. 一种抗旋转的全息水印算法[J]. 包装工程,2013,34(19):95—100.
ZHOU Zhong-yuan, SUN Liu-jie, TANG Bo, et al. An Anti-rotation Holographic Watermarking Algorithm[J]. Packaging Engineering,2013,34(19):95—100.

[14] DO M N, VETTERLI M. Contourlets: A New Directional Multiresolution Image Representation[C]// Signals, Systems and Computers,2002,1:497—501.

[15] DO M N, VETTERLI M. The Contourlet Transform: An Efficient Directional Multiresolution Image Representation[J]. IEEE Trans. Image Process. 2005,14(12):2091—2106.