# 抗攻击的图像水印技术综述

李云亚, 李冬森, 付　蓉

（江苏金盾检测技术有限公司, 江苏　南京　210042）

**摘　要:** 文章简要地探讨了抗各种攻击图像水印技术, 包括一些基本概念: 水印概念、水印的嵌入和提取过程、水印的分类和衡量水印的基本指标。文章把几何攻击分为图像变换攻击和剪裁攻击, 图像处理攻击分为噪声攻击、滤波攻击、JPEG攻击和打印扫描拍照攻击。此外, 文章探讨了图像水印抗击几何攻击和抗图像处理攻击从传统到最新的技术和方法。

**关键词:** 攻击; 抗攻击; 图像水印

## 0　引言

21世纪是一个数据大爆发的年代, 在《数据时代2025》报告中显示, 全球每天数据量将达到491EB（EB: 百亿亿字节）。从远古时代的纸素书到今天的微信传输, 智能终端的快速发展, 让互联网传输变得司空见惯。互联网带来方便的同时, 也面临着巨大的安全隐患, 数字作品在互联网上被恶意抄袭和篡改。数字水印技术是主动保护数字作品的有效手段。数字水印技术可以在有争议的数字作品上为鉴别真伪提供合理依据。图像作为数字作品中占有一席之地, 从绘画作品到设计再到产品商标, 对于这些图像的保护就是对图像背后的商家或是作者版权的保护。

1993年Tirkel等[1]人发表论文“Electronic Watermark”, 这篇论文是图像水印技术的开端, 首次提出在数字图像的空间域上修改最低有效位来隐藏信息（Least Significant Bit, LSB）。这种方法在数字图像上极易实现而且操作也十分简单, 但是鲁棒性很差, 基本不具备抗攻击的性能。1995年Cos等[2]人提出一个经典的图像水印算法, 去修改频域图像中的DCT系数从而将水印信息隐藏到载体图像中, 但是这一方法是非盲水印需要提供原始未经修改的载体图像才能在解码过程中得到水印信息, 在实际操作中我们获取到带有水印的图像很难找到没有水印的原始载体图像, 这对水印的提取带来很大的挑战。1998年Ruanaidh等人提出一种经典的抗RST（旋转, 缩放, 平移）几何攻击的图像水印算法, 将水印信息嵌入RST几何攻击不会改变的Fourier-Mellin域中, 从而使含有水印的图像具备抗RST几何攻击的特性, 此算法的提出为后续研究者研究抗几何攻击的图像水印技术提供了新的思路。2000年以后很多技术的提出是根据离散余弦变换、离散小波变换和傅里叶变换这些技术进行改进的, 这样的改进也使图像水印可以更好地抵抗几何攻击, 而且含有水印的载体图像的质量也随之提高。但是随着信息时代的快速发展, 对图像水印的攻击也不仅仅局限于几何攻击。抗JPEG压缩攻击的提出很快就得到了回应, 原始的基于离散小波变换和傅里叶变换的图像水印技术, 不仅具有抗几何变换的性能也具备抗JPEG压缩攻击的性能。2010年Wang等人提出基于特征的半色调图像水印方法, 这个方法可以使图像水印抵抗打印扫描攻击。随着手机逐渐成为大众生活的必需品, 从之前用扫描机来扫描图像到现在用手机去拍照图像, 不管是手机拍照图像还是用扫描机扫描图像对图像的攻击不尽相同。2020年Matthew Tancik等[3]人提出用利用深度学习网络研究图像水印问题, 提出了一种抗打印、拍照、旋转、JPEG压缩等攻击的图像水印技术。

自水印第一次被提出, 水印的鲁棒性就成为大众关注的焦点, 它是否具备抗攻击的性能成为评价水印技术的一个指标。攻击和抗攻击是一对矛和盾, 研究者不断研究更鲁棒的数字水印算法, 攻击者不断修改攻击的方式, 二者博弈至今。

本文首先在图像水印的基础上, 探究抗攻击的图像水印技术, 其次将攻击类型分类, 传统的分类方式是从攻击者角度将水印攻击分为: 噪声攻击、同步攻击、欺骗攻击和共谋攻击。本文从研究者出发将抗攻击的图像水印技术分为: （1）抗几何攻击的图像水印技术, 从21世纪开始图像水印技术就围绕着抗击几何攻击展开, 使水印图像不断具有抗旋转、拉伸、缩放和剪裁等性能。（2）抗图像处理攻击的图像水印技术, 对图像的锐化, 高斯模糊, 亮度的增加和降低, 对图像进行JPEG压缩, 这些对图像的处理, 都会导致水印无法提出。对图像打印扫描和拍照可以看成是对图像加了噪声或者滤波, 同样会导致水印无法正常提出。最后, 本文总结

**作者简介:** 李云亚（1977— ）, 男, 江苏南京人, 高级测评师, CISP, 高级项目管理师, 商用密码测评师, CIIPT, 学士; 研究方向: 网络应用与安全。

目前抗攻击的图像水印现状和存在的问题并展望未来的发展方向。

## 1 相关知识

### 1.1 数字水印基本概念

图像水印是信息隐藏的一个技术分支，通过将水印信息隐藏到载体数字媒介中，载体数字媒介可以是图像、文本、音频或是视频。通过水印信息对数字作品进行版权保护，在发生版权纠纷时，提供有利的证据支持，为数字作品的版权归属和防止侵权提供有效的方式方法。数字水印技术可以通过加密算法将公司的商标，创作者的个人标签和产品标识等嵌入数字作品，必要时通过解密算法可以测试数字作品中是否有水印信息，通过这种方式对作品进行版权保护。

### 1.2 数字水印嵌入和提取过程

数字水印技术可以分为水印的嵌入过程和水印的提取过程。水印的嵌入算法设计是依据水印的3个主要特征不可感知性、鲁棒性和水印容量所设计的，不断优化嵌入算法使它在鲁棒性和不可感知性提高的前提下尽最大可能提高水印的容量。提取算法要根据水印类型盲水印和非盲水印来设计，但实际需求中盲水印的实际价值要远远大于非盲水印的需求。数字水印的嵌入和提取过程如图1所示。

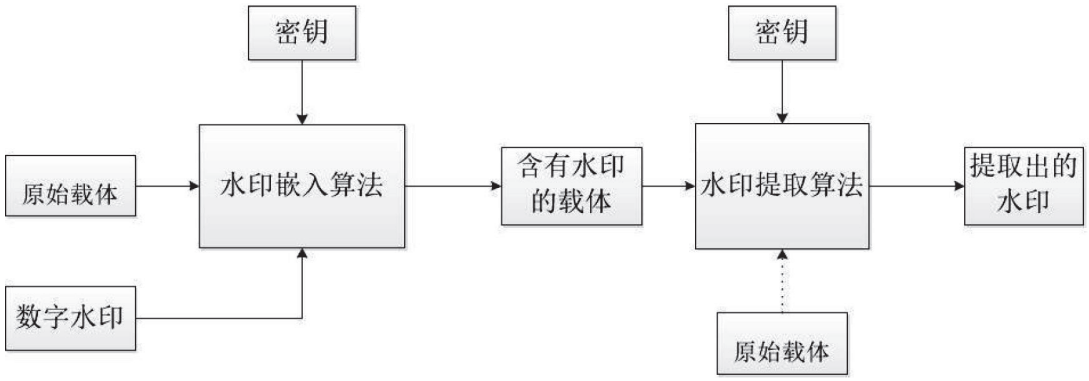


**图1 数字水印的嵌入和提取过程**

### 1.3 数字水印的分类

对于数字水印可以按照多种方法分类如下。

按照原始载体的类型进行分类，可以分为图像水印、音频水印、文本水印和视频水印。现实生活中总会有来自不同数字载体的水印需求，这些数字产品的所有者都不希望在网络传播中受到侵权或是恶意破坏，当数字产品遭到侵权或者是恶意破坏时，数字产品的版权所有者因为数字产品上含有水印可以拿起法律的武器保护自己的权益。2020年张新鹏等人提出神经网络水印，目的是给那些发布在网络上或是在云端的网络框架加水印保护其版权。随着信息技术的蓬勃发展和版权意识的不断提高，需要用水印保护的数字产品远不止是原来的图像、文本、视频和音频，将来还会有更多类型的数字产品作为原始载体需要加水印保护版权。

按照水印的嵌入位置进行分类，分为空域水印和频域水印。空域水印是将水印信息直接隐藏在空域信号中。典型的方法有最低有效位法、像素直方图法和利用图像信息之间的周期性和相关性法。空域水印操作相对简单但是鲁棒性较差。频域水印是将水印信息隐藏在原始载体某种变换域的变换系数上。经典方法有离散余弦变换（DCT）、离散小波变换(DWT)、傅里叶变换（DFT）和奇异值分解（SVD）。频域水印与空域水印相比较具有很强的抗攻击的特性。

按照人类的感知进行分类，分为可见水印和不可见水印。将不可见水印按照水印的抗攻击的能力又可分为脆弱水印、半脆弱水印和鲁棒水印。脆弱水印对攻击的感知能力极强，用于原始载体的数据内容完整性检测和载体信息是否被破坏都可通过脆弱水印检测出来。半脆弱水印是指对特定的操作具有鲁棒性，而对其他操作不具备鲁棒性。鲁棒水印可以抵抗大多数攻击，可以用于对原始载体的版权保护。

按照水印的提取方式进行分类，分为非盲水印和盲水印。非盲水印在水印提取时，需要提供原始的载体信息才能将水印提取出来。相反，盲水印在水印提取时仅需要提供含水印的载体即可得到水印信息。由于在实际应用中载体信息很难获取，所以相比较而言，盲水印在实际应用中更为广泛，实用价值更高。

### 1.4 数字水印衡量的基本指标

#### 1.4.1 不可感知性

将水印嵌入原始载体，势必会引起原始载体发生一些变化，原始载体的变化越不容易引起怀疑越不可感知，说明水印嵌入是成功的，因此引入峰值信噪比（PSNR）的概念计算载体在嵌入前后的变化从而去衡量含有数字水印的原始载体的不可感知性。通常情况下，水印嵌入前后峰值信噪比低于30 dB，说明水印的不可感知性比较差，相反如果水印嵌入前后峰值信噪比高于40 dB，肉眼很难看出水印嵌入前后的差别，说明水印的不可感知性较好。

1.4.2 鲁棒性

数字水印的鲁棒性可以用来衡量水印抗攻击性的能力。数字水印在网络或者是现实生活中都很容易受到各种各样的攻击，比如典型的几何攻击、压缩攻击、打印攻击等等。为了评价水印的鲁棒性，引入误码率（Bit Error Rate, BER）和归一化相关系数（Normalized Cross Correlation）两个评价指标。一般地，误码率越低表示提取出的水印信息正确率就越高，归一化相关系数越低，则说明提取出的水印信息与原始水印信息相差很大，从而一个鲁棒性强的算法它的误码率越低越好，相反，归一化相关系数越高越好。

1.4.3 嵌入容量

衡量数字水印的另一个指标就是水印的嵌入容量，既要考虑水印的不可感知性，又要考虑水印的鲁棒性，在这两个性能的前提下还需要尽可能地提高水印的嵌入容量。水印的基本信息如图2所示。

## 2 抗几何攻击的图像水印技术

### 2.1 抗图像变换攻击的图像水印

#### 2.1.1 抗图像变换攻击的图像水印

在图像变换中最常见的就是RST（旋转、缩放、平移）几何变形，也是研究者在图像水印中首先考虑的攻击形式，在对图像进行旋转、缩放或者是平移操作后，水印是否能被顺利提取出来。

1998年Ruanaidh等人提出一种经典的抗RST几何攻击的图像水印算法，将水印信息嵌入RST几何攻击不会改变

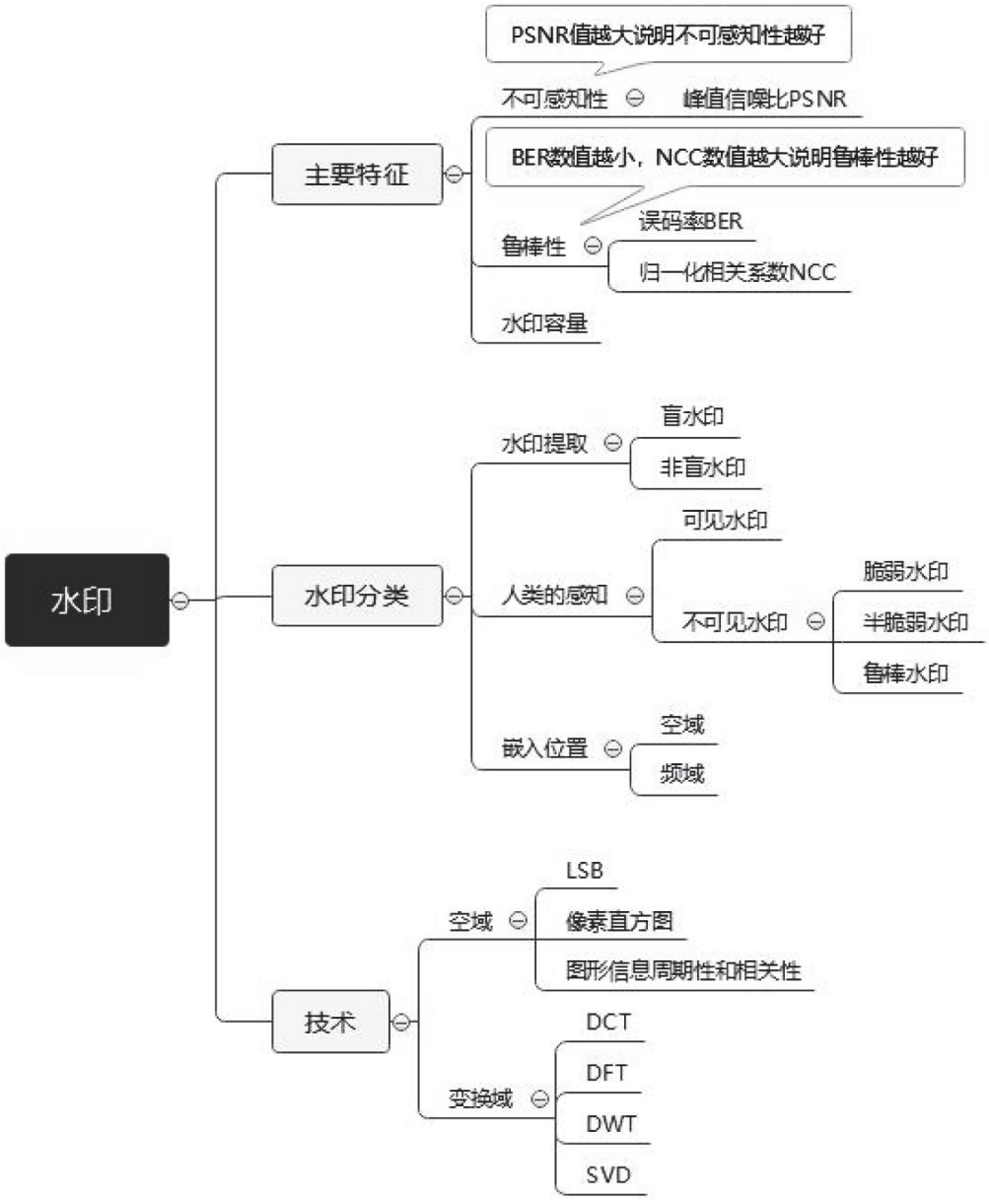


图2 水印的基本信息

的Fourier-Mellin域中，从而使含有水印的图像具备抗RST几何攻击的特性。2001年Lin等[4]人对上述方法进行改进，利用Fourier-Mellin变换来对抗缩放和平移攻击，但是算法仍然存在图像质量下降严重的缺点。基于Fourier -Mellion变换（FMT）的水印算法，利用傅立叶梅林变换不变量的理论，可用于产生抗旋转、缩放和平移的水印。将水印嵌入Fourier-Mellin域。

2003年Zheng等[5]提出基于对数极性映射和相位相关的RST不变数字图像水印。基于对数极坐标映射（LPM）和相位相关性，水印嵌入原始图像的傅立叶幅值频谱的LPM中，这样水印不便于旋转、缩放和平移（RST）。并使用原始图像的LPM和水印图像的LPM之间的相位相关性来计算LPM域中水印位置的位移，通过避免计算逆对数极坐标映射（ILPM）来保持图像质量，并通过使用相位相关来避免详尽搜索，从而为未加水印的图像生成较小的相关系数。

2008年Xiang等[6]人提出基于低频域统计特征的不变图像水印，对抵抗RST攻击和随机弯曲攻击都具备良好的性能。基于低频域统计特征，通过在图像的高斯滤波低频分量中使用两个统计特征（直方图形状和均值）来提出图像水印算法。

2009年Zheng等[7]人继续提出一种数学建模与水印过程分析的RST不变图像水印算法，利用数学方法来分析水印问题并且取得了很好的效果。数学建模与水印过程分析的RST不变图像水印算法，基于旋转不变特征和图像归一化的旋转定标不变图像水印方案。建立了基于混合广义高斯分布的近似图像的数学模型，可以促进水印过程的分析。使用基于最大后验概率的图像分割，将封面图像分割为几个均匀的区域。每个区域都可以由广义的高斯分布表示，这对于从数学上分析水印过程至关重要。从分割区域中提取旋转不变特征，并将其选择为参考点。以特征点为中心的子区域用于水印嵌入和提取。图像归一化应用于子区域以实现缩放不变性。同时，在建立的数学模型的基础上，对水印的嵌入和提取方案进行了数学分析。使用噪声可见性函数自适应地调整水印嵌入强度，并通过数学分析错误概率，建立了不可感知性和鲁棒性之间的数学关系。

2011年Lin等[8]人提出一种采用动态规划的同步水印算法，对RST攻击的抵抗方面变现突出。2012年Owalla 等人提出提出了一种既能抵抗信号处理又能抵抗几何攻击的鲁棒数字图像水印方案。

2.1.2 图像变换攻击类型分析

RST攻击包括旋转、平移和缩放，但是随着计算机技术的日益精进演变出多种多样复杂的图像变换，不仅仅局限于旋转、平移和缩放，还有放射变换、透视变换和随机弯曲等等。该算法将水印以扩频格式和用于压缩图像的矢量量化技术嵌入离散余弦变换域。旋转、缩放和平移攻击后水印的恢复是通过使用基于Harris角检测器的特征点获得Delaunay细分来完成的，该细分用于逆转攻击。 在RST攻击导致在图像上形成大量暗区的情况下，某些参考特征点会丢失，恢复的水印会很差或完全丢失。 此算法通过对镶嵌中选定三角形的平均值进行估算，来估计RST攻击的过程，从而达到抗RST攻击的特性。

2.2 抗裁剪攻击的图像水印

剪裁攻击一般会对图像其中一部分剪裁或者随机剪裁掉图像其中一部分，在剩余图像中依然可以提出水印信息。早在1999年 Hsu等人通过选择性地修改图像的中频部分将水印信息嵌入图像，从而达到抗裁剪的特性。在2001年上文中提到的Lin等人利用Fourier-Mellin变换来对抗RST攻击，这个算法对裁剪攻击依旧有很好的效果。

2003年Tang[9]提出一种结合图像特征提取和图像归一化的鲁棒数字图像水印方案，它不仅可以抵抗裁剪攻击还可以抵抗几何失真和滤波攻击。它采用MHW（Mexican hat wavelet）尺度交互作用的特征提取方法。

在上文提到的2008年Xiang 等人提出基于低频域统计特征的不变图像水印，也可以抵抗裁剪攻击。

2015年Zong等[10]人提出基于鲁棒直方图形状的图像水印方法，该方法在嵌入过程中通过高斯低通滤波器对主机图像进行预处理。然后，使用密钥来随机选择多个灰度级，并且构建关于这些所选灰度级的滤波图像的直方图。此后，引入直方图形状相关索引以选择像素数最多的像素组，并在选择的像素组和未选择的像素组之间建立安全带。提出了一种水印嵌入方案，以将水印插入所选的像素组中。直方图形状相关的索引和安全带的使用导致了良好的鲁棒性。此外，在嵌入方案中还使用了新颖的高频分量修改机制来进一步提高鲁棒性。在解码端，基于可用密钥，识别出加水印的像素组，并从中提取水印。这种方法可以抗击裁剪攻击和随机弯曲攻击。

2018年 Loan等[11]人提出了一种适用于灰度和彩色图像的基于混沌加密的盲数字图像水印技术，可以抵抗裁剪攻击。论文提出了一种适用于灰度和彩色图像的基于混沌加密的盲数字图像水印技术。在将水印嵌入宿主图像之前，先使用离散余弦变换（DCT）。在应用DCT之前，主机图像被划分为8×8个非重叠块，并且通过修改相邻块的DCT系数之间的差来嵌入水印位。

2020年Meenpal等[12]人提出基于对偶树复小波变换的数字水印技术，使用对偶树复小波变换的新技术进行数字水印的方法。通过将RGB图像转换为YUV通道，将水印嵌入色

度级别中。这样做是因为色度水平对人类视觉系统的感知较小，使用预定义的低通滤波器会生成感知蒙版，这会增加水印的不可感知性。同时也具备抗旋转和裁剪的特征。

## 3 抗图像处理攻击的图像水印技术

### 3.1 抗滤波和噪声攻击的图像水印

滤波攻击分为高斯滤波和中值滤波两种攻击，但是由于滤波攻击并非主流研究方向，仅找到上文提到的2003年Tang提出一种结合图像特征提取和图像归一化的鲁棒数字图像水印方案，可以抵抗几何失真和滤波攻击。对图像加噪声会对水印的提取造成一定的干扰，一般噪声攻击对图像加入高斯噪声。基于DFT，DCT，DWT和SVD等这样的技术和算法，都可以抵御噪声攻击。近几年迅猛发展的深度神经网络架构，基于深度学习的水印技术对噪声攻击的抵抗效果也良好，这里不引入论文赘述。

### 3.2 抗JPEG攻击的图像水印技术

随着互联网通信的快速发展，在互联网中传输图片信息也越来越普遍，在传输过程中图片会被压缩，最常见的压缩是JPEG压缩。JPEG压缩的过程图如图3所示。

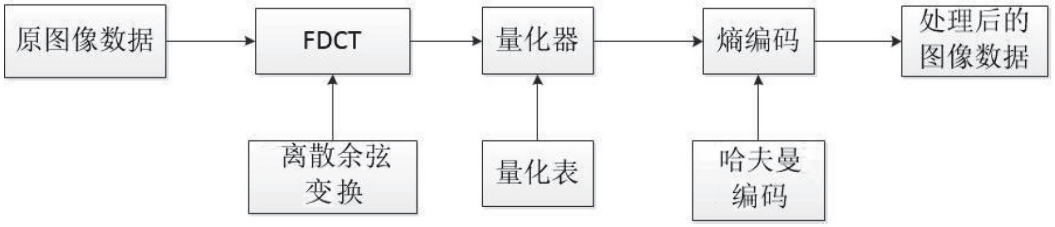


图3 JPEG压缩的过程

2003年Afzel[13]提出了一种新的半盲水印修正离散余弦的JPEG图像技术转换（MDCT）。但是半盲水印在水印提取时需要提供原载体图像，由于原图像在实际生活中很难获得，所以这个技术有很大的局限性。

2010年Wang Xiangyang提出一种基于特征的半色调图像数字水印方法。利用多尺度从主体半色调图像中提取特征点哈里斯-拉普拉斯探测器和局部特征区域（LFR）是根据特征构造的规模理论。其次，对LFR执行离散傅里叶变换（DFT），然后进行嵌入并根据幅度谱信息自适应地选择位置（DFT系数）。最后，通过量化所选水印的幅度，将数字水印嵌入到LFR中DFT系数,这种方法就解决了上述的半盲水印的局限。

### 3.3 抗打印扫描拍照攻击的图像水印

之前研究的内容就是对互联网上的图像水印进行各种攻击，但是现实世界中对水印也有很多的应用，包括将含有水印的图像从打印机中打印出来，和在用扫描机将其扫描回计算机中提取出水印。手机渐渐成为生活的必备品，用手机拍照上传至互联网中来代替以前用扫描机将图片扫描到计算机中，从打印到扫描或者是拍照都对含有水印的图像是一种攻击。

2004年牛少彰等人提出抗打印扫描的数字水印鲁棒性，文章中发现打印之后的图像和打印之前的图像DCT系数是几乎不变的，算法将分块的图像进行DCT变换，然后将所有的DCT系数按照位置进行特殊的分组，通过一定的嵌入强度调整每组中DCT系数正负号的个数来实现水印信息的嵌入。

2005年Yu等人提出基于DFT的水印技术，将水印信息嵌入到幅值中得到含水印的幅值谱，通过DFT逆变换得到含水印图像。

2010年洛锦提出一种基于CIELab色彩空间分块DCT变换域的彩色图像盲水印算法，算法中通过对打印扫描前后DCT系数的统计,根据对DCT系数的详细分析,将水印信息嵌入到L分量DCT系数的中频区域，每3个DCT系数为一组嵌入一位水印信息，通过调节中间值到两端值之间的距离进行水印的嵌入。算法很好地利用了DCT系数的冗余性，大大提高了水印的容量。

2014年Mirza提出一种基于混合域的抗打印扫描数字水印技术，将水印信息及嵌入空域中也嵌入到频域中，这样提高了数字水印的鲁棒性。在此期间很多研究者基于DCT和DFT技术做出改进和创新提出抗打印扫描的图像水印技术。

2020年Tancik等[3]人提出基于深度学习的图像水印技术，利用神经网络架构对水印进行嵌入和提取。在抗打印扫描和拍照方面有着良好的效果。但是也有很多不足之处，比如水印容量较小，并且图像质量不如传统方法嵌入水印的图像质量好。

## 4 结语

本文对抗各种攻击的图像水印进行归纳和总结，将攻击类型分类，总结各种攻击类型所应用的图像水印技术，在归纳中发现很多图像水印算法抗某种攻击的能力不是单独存在的，往往一个图像水印算法可以抵抗多种攻击类型。但是随着科学技术的不断发展和进步，提出各种各样

的新的攻击形式，这也对图像水印的研究者带来新的挑战。深度学习和神经网络的普遍应用，让图像水印有了与传统图像水印截然不同的嵌入和提取方法，基于深度学习的图像水印抗攻击的性能良好，但是在水印容量、图像质量和算法复杂程度上还有改进的空间，给研究人员带来了新的研究方向。

[参考文献]

[1]TIRKEL A Z, RANKIN G A, SCHYNDEL R M V, et al.Osborne.electronic watermark[J].Digital Image Computing Technology and Applications, 1993(1): 666-673.

[2]COX I J, KILLIAN J, LEIGHTON T.et al.Secure spread spectrum watermarking for multimedia[J].IEEE Trans on Image Processing, 1997(12): 1673-1687.

[3]TANCIK M, MILDENHALL B, NG R.Invisible hyperlinks in physical photographs[C].Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Stegastamp: 2020.

[4]LIN C Y, WU M, BLOOM J A, et al.Rotation, scale, and translation resilient watermaking for images[J].IEEE Transactions on Image Processing, 2001(5): 767-782.

[5]ZHENG D, MEMBER S, ZHAO J, et al.RST-invariant digital image watermarking based on log-polar mapping and phase correlation[J].IEEE Transactions on Circuits & Systems for Video Technology, 2003(8): 753-765.

[6]XIANG S, KIM H J, HUANG J, et al.Invariant image watermarking based on statistical features in the low-frequency domain[J].IEEE Transactions on Circuits & Systems for Video Technology, 2008(6): 777-790.

[7]ZHENG D, WANG S, ZHAO J.RST invariant image watermarking algorithm with mathematical modeling and analysis of the watermarking processes[M].Piscataway: IEEE Press, 2009.

[8]LIN Y T, HUANG C Y, LEE G C.Rotation, scaling, and translation resilient watermarking for images[J].Image Processing Iet, 2011(4): 328-340.

[9]TANG C W, HANG H M.A feature-based robust digital image watermarking scheme[J].IEEE Transactions on Signal Processing, 2003(4): 950-959.

[10]ZONG T, XIANG Y, NATGUNANATHAN I, et al.Robust histogram shape-based method for image watermarking[J].IEEE Transactions on Circuits&Systems for Video Technology, 2015(5): 717-729.

[11]LOAN N A, HURRAH N N, PARAH S A, et al.Secure and robust digital image watermarking using coefficient differencing and chaotic encryption[M].Piscataway: IEEE Access, 2018.

[12]MEENPAL A, MAJUMDER S, BALAKRISHNAN A.Digital watermarking technique using dual tree complex wavelet transform[M].Piscataway: IEEE Access, 2020.

[13]NOORE A, VIRGINIA W.An improved digital watermarking technique for protecting[C]//IEEE International Conference on Consumer Electronics, Las Vegas, 2003.

（编辑　王雪芬）

# Summary of anti-attack image watermarking technology

Li Yunya, Li Dongsen, Fu Rong

（Jiangsu Golden Shield Detection Technology Co., Ltd.,Nanjing 210042, China）

Abstract: This article briefly discusses image watermarking techniques against various attacks. This article includes some basic concepts: watermark concept, watermark embedding and extraction process, watermark classification and basic indicators to measure watermark. In this paper, geometric attacks are divided into image transformation attacks and cropping attacks, and image processing attacks are divided into noise attacks, filtering attacks, JPEG attacks, and print scanning and photographing attacks. In addition, discuss the traditional to the latest technologies and methods for image watermarking against geometric attacks and image processing attacks.

Key words: attack; anti-attack; image watermark