

基于图像攻击的隐藏信息盲检测技术

毛家发^{1),2)} 林家骏²⁾ 戴 蒙³⁾

¹⁾(上饶师范学院数学与计算机系 江西 上饶 334001)

²⁾(华东理工大学信息与控制学院 上海 200237)

³⁾(上海应用技术学院计算机科学与工程系 上海 200233)

摘 要 隐写分析技术是网络信息安全的重要组成部分,它的最根本问题是判断数字载体是否携带秘密信息.文中提出了基于攻击的隐藏信息盲检测方式,提取空域图像、DCT域图像一个二维特征向量,通过非线性神经网络分类器来判别图像是否含有秘密信息,并建立了7000多幅图像库(包括掩密图像与干净图像)进行了可行性、多样性仿真实验,取得了以下效果:(1)平均检测率90.085%(阳性检测率与阴性检测率的总平均);(2)不受限于隐写方案;(3)能够检测低嵌入率的掩密图像;(4)能够检测出经平滑、锐化、缩小、剪切和再压缩处理后的干净图像.

关键词 隐写术;隐写分析;数字水印;数据隐藏

中图法分类号 TP391 DOI号: 10.3724/SP.J.1016.2009.00318

An Attacked Image Based Hidden Messages Blind Detect Technique

MAO Jia-Fa^{1),2)} LIN Jia-Jun²⁾ DAI Meng³⁾

¹⁾(Department of Mathematics and Computer Science, Shangrao Normal University, Shangrao, Jiangxi 334001)

²⁾(College of Information Science & Engineering, East China University of Science & Technology, Shanghai 200237)

³⁾Computer Science and Information Engineering Department, Shanghai Institute of Technology, Shanghai 200233)

Abstract Steganalysis technique is an important part of the security of internet information. It's essential problem is to estimate if the digital carrier takes the secret messages. This paper brings forward the blind detect technique which is based on the hidden messages for aggressive. A 2D feature vector in the spatial and DCT domains is extracted to blindly determine the existence of hidden messages in an image. For evaluation, a database composed of over 7000 cover and stego images (generated by using ten different embedding schemes) was established. Based on the database, simulation experiments were conducted to prove the feasibility and diversity of the proposed system. The proposed system consists of following: (1) The average detection rate is about 90.085%, the total average for electropositive detection rate and negative detection rate; (2) It is not limited to the detection of a particular steganographic scheme; (3) It has the capability to detect the stego images with an lower embedding rate; (4) The test of plain images which is smoothing, sharpening, image size zoomed out, cutting and recompressing can be inspected.

Keywords steganaography; steganalysis; digital watermarking; data hidden

收稿日期:2006-10-25;最终修改稿收到日期:2008-12-16. 本课题得到江西省教育厅科技基金项目(Gjj08462)、上海市教育发展基金会晨光计划(2008CGB21)资助. 毛家发,男,1970年生,博士研究生,副教授,主要研究方向为数字水印、隐写分析、模式识别、图像与信号处理. E-mail: maojiafa@21cn.com. 林家骏,男,1948年生,博士,教授,博士生导师,主要研究领域为模式识别、智能控制、信号检测、数字水印和隐写分析. 戴 蒙,女,1979年生,博士,主要研究方向为模式识别、智能控制、隐写分析和图像处理.

1 引言

众所周知, 隐写术与数字水印是信息隐藏技术的两个主要应用, 尽管在一些方面存在不同, 但它们在数据嵌入与提取方式方面有许多共同之处。如鲁棒性、嵌入容量等。在特定的场景里, 水印信息作为认证多媒体作品版权归属的一个重要依据。因此, 隐写分析在信息隐藏与数字水印方面有着广泛应用前景。

本文仅仅是做图像的隐藏信息检测(不考虑声音和视频的隐藏信息检测)。对于隐写分析者而言, 隐写方式是未知的。因此, 我们只关心图像是否含有秘密信息, 不关心其是用哪种隐写术嵌入的, 也就是说是一个盲检测问题。我们提出基于攻击特征算法(空域与 DCT 域各一个二维特征向量), 利用非线性神经网络分类器来实现隐藏信息盲检测, 并用 Matlab 仿真实验来进行算法评估。

本文第 2 节详细介绍隐写术与隐写分析当前现状; 第 3 节是基于攻击的特征提取; 第 4 节介绍神经网络分类器; 第 5 节是仿真实验; 第 6 节给出总结与展望。

2 隐写术与隐写分析

图 1 表示的是一个大多数隐藏信息过程, 如果一个发送者想借助于互联网发送秘密信息, 他将选择一种媒体(原始信息), 将秘密信息经加密后嵌入到原始信息中, 得到掩密信息, 而掩密信息与原始信息在视觉上几乎没有差别。而接受者受到掩密信息的, 经提取解密后得到秘密信息和原始媒体信息。

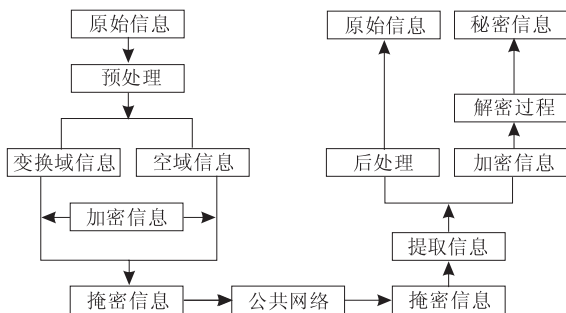


图 1 隐藏信息过程

传统的图像信息隐藏可以分为两类: 空间域信息隐藏和变换域(如 DCT 变换域、小波变换域)信息隐藏^[3]。对于空间域方式, 在大多数文献中, 直接加入^[4-6]到原始图像信息中或直接嵌入到原始图像

信息中的最不重要位(LSB)^[7-8]上。经这些技术处理过的掩密图像, 在内眼上与原始图像几乎没有差别, 尽管如此, 计算机数字识别系统还是能够分析到这微小的差别。一些工具, 如 S-TOOLS、Imagehide、StegoDos、EzStego, 都是基于空间域隐写方式的^[3,9]。

针对上面所提到的空间域隐写算法, 目前已经有一些相对应的隐写分析算法^[10-12]。Fridrich 等^[11]提出了基于位平面隐写分析算法, 他们认为图像相邻的像素或多或少存在相关性, 而且 LSB 可以从其他 7 个位平面估计出。当嵌入了隐藏信息后, 这种相关性会减弱; Kong 等^[13]提出了利用统计滤波器评估图像的复杂性, 他们认为含有秘密图像的 LSB 比未掩密图像的 LSB 更具随机性; Cabanillas 等^[10]假设 LSB 服从高斯分布模型, 提出最大后置检测器(MAP)并估算最大嵌入容量。遗憾的是, 在他们的论文中既没有给出 MAP 检测器的实际应用也没有给出具体的实验结果, 而且他们的理论分析严格地受限于嵌入后 LSB 必需服从高斯分布模型; Maes^[14]认为强度恒定的空域隐藏算法(如传统的 LSB 隐写方式), 会导致图像灰度直方图出现一个局部峰值, 而这个局部峰值能够进一步分离成两个小峰值, 这种现象被称为孪生双峰值效应(twin-peak effect), 这种孪生双峰值效应也存在于隐藏强度不固定的隐写技术, 只要嵌入信息是统计独立的。但是, 他并没有提到怎样在数量上去度量孪生双峰值; Avcibas 等^[12]提出任何图像质量经平滑或低通滤波后都会退化, 这种退化量与图像是否含有秘密信息有一定的关联关系。因此, 他们利用回归分析来度量图像质量的退化量进行隐写分析。

信息隐藏也能在变换域进行, 如 DCT 变换域^[4,15-20]或小波变换域^[18,21]等。不管 DCT 变换域还是小波变换域, 重要的是隐藏了秘密信息或水印的变换域系数对肉眼来说都应具有不可见性。例如, Cheng 等^[20]提出了一种加性方式将秘密信息隐藏在 DCT 或小波变换域中; Wu 等^[19]提出了二级数据嵌入方式, 主要是将秘密信息附加到展谱和频谱片中去, 用来复制控制、访问存取控制、鲁棒注释和基于内容鉴定等。像目前存在的隐写工具如 J-Steg 和 Outguess 等, 都属于这种技术范畴^[3,22]。

在一些隐写分析^[11,23-24]中提出了基于 DCT 变换域隐藏信息检测方法, Manikopoulos 等^[23]提出了通过计算掩密图像的 DCT 系数概率密度函数(PDF)和干净参考图像的 PDF, 把 PDF 当作一个特

征输入到基于训练的二层神经网络进行分类. 在他们的论文中, 干净参考图像的 PDF 是所有原始干净图像库中图像 PDF 的平均, 但是, 这些原始干净参考图像的代表性是成问题的. Fridrich 等^[11] 认为 JPEG 格式的图像隐藏了信息后其像素块将饱和 (即至少一个像素灰度值是 0 或 255). 如果图像块找不到饱和块, 则认为此图像不含有秘密信息. 在干净图像与掩密图像之间, Chandramouli^[24] 把普通的隐写术看成是一个线性转换, 在一幅掩密图像经过线性转换后至少获得两个副本, 这就类似于独立向量分析 (ICA) 技术来分离盲资源问题. Harmsen 等^[25] 认为在一幅图像嵌入秘密信息相当于图像加入加性噪声, 掩密图像的直方图特征函数比干净图像的直方图特征函数更聚中, 通过这种特点他们把掩密图像从干净图像中分离出来.

Lei 等^[26] 提出了基于空间域和 DCT 变换域隐藏信息检测方法. 首先在空间域和 DCT 变换域分别提取一个特征构成一个二维特征向量, 然后将二维特征向量输入三层前馈非线性神经网络分类器把掩密图像与干净图像分开. 在空间域中提取图像的梯度能量, 他们认为图像在嵌入信息前的梯度能量小于嵌入信息后的梯度能量, 并在理论上加以证明; 在 DCT 变换域中估算 Laplacian 参数方差作为一个特征, 他们假设图像的 DCT 系数中的 AC 系数是服从 Laplacian 分布的, 并从理论上推导出掩密图像的 Laplacian 参数方差比未掩密图像的小.

我们看到, 大多数隐写分析法都是基于一个特写的操作域设计的, 或者是针对特定的隐写算法的. 至今为止, 一个通用的隐藏分析系统很难找到.

3 基于攻击的特征提取

文献^[22] 提出应用统计特性来隐藏信息检测, 如 LSB 位平面的随机性、图像像素间灰度级的改变、变换域系数的变化等. 一般来说, 不存在单一特征能够有效地将掩密图解与原始图像很好地分开. 从不同领域提取多个特征构成特征向量的效果比单特征效果更好. 考虑到目前隐写算法都是在两个范畴嵌入的: 图像的空域 (如 stools, imagehide, ipv, itp 等工具) 和 DCT 变换域 (如 F5, Jsteg, jpide, out-guess 等工具). 因此, 我们系统在每个范畴提取两个特征构成一个二维特征向量. 对于空域格式图像的两个特征, 是由梯度能量特征和经平滑攻击后梯度能量比特征组成; 对于 JPEG 格式图像的两个特

征, 是由图像能量特征和经再压缩攻击后的图像能量差特征构成.

3.1 空域图像特征: 梯度能量和梯度能量比

3.1.1 梯度能量特征

参考文献^[4-5, 7, 14, 26], 一个通用隐藏信息公式定义如下:

$$I^h = I^o + \beta \cdot w \quad \text{或} \quad I^h = I^o (1 + \beta \cdot w) \quad (1)$$

这里的 I^h 和 I^o 分别表示掩密图像和原始图像的像素值, w 表示传输的信息, β 是个尺度因子或称嵌入强度. 当 LSB 嵌入时, 式 (1) 的前部分中的 $\beta \cdot w$ 的值是 -1、0 和 1; 如果 β 很大时, 嵌入的信息会使图像失真, 也就是说人的肉眼能看出, 如明水印信息, 其强度可以达到 256.

下面考虑一维序列的梯度能量情况. 假设一维序列 $I(n)$, 我们定义梯度 $r_I(n)$ 是相邻两个序列点的差:

$$r_I(n) = I(n) - I(n-1) \quad (2)$$

序列 $I(n)$ 的梯度能量 GE 定义为

$$\begin{aligned} GE &= \frac{1}{N-1} \sum_{n=2}^N (r_I(n))^2 \\ &= \frac{1}{N-1} \sum_{n=2}^N (I(n) - I(n-1))^2 \end{aligned} \quad (3)$$

这里的 N 是序列 $I(n)$ 的长度.

定理 1. 掩密信号序列 I^h 的梯度能量均值 $E[GE_I^h]$ 大于等于原始信号序列 I^o 的梯度能量均值 $E[GE_I^o]$ ^[26].

证明. 由式 (3) 知, 原始信号序列 I^o 的梯度能量均值为

$$E(GE_I^o) = \frac{1}{N-1} \sum_{n=2}^N E[(r_I^o(n))^2] \quad (4)$$

掩密信号序列的梯度为

$$\begin{aligned} r_I^h(n) &= I^h(n) - I^h(n-1) \\ &= (I^o(n) + q(n)) - (I^o(n-1) + q(n-1)) \\ &= I^o(n) - I^o(n-1) + q(n) - q(n-1) \\ &= r_I^o(n) + r_q(n) \end{aligned} \quad (5)$$

这里的 $q(n)$ 表示扰动序列, 如 $q = \beta \cdot w$ 或 $q = I^o \cdot \beta \cdot w$, $r_q(n) = q(n) - q(n-1)$, 那么, 嵌入信息后的梯度能量均值为

$$\begin{aligned} E[GE_I^h] &= E\left[\frac{1}{N-1} \sum_n (r_I^h(n))^2\right] \\ &= \frac{1}{N-1} \sum_n E[(r_I^o(n) + r_q(n))^2] \\ &= \frac{1}{N-1} \sum_n E[(r_I^o(n))^2 + \\ &\quad 2 \cdot r_I^o(n) \cdot r_q(n) + (r_q(n))^2] \end{aligned}$$

$$= \frac{1}{N-1} \sum_n \{E[(r_i^o(n))^2] + 2E[r_i^o(n) \cdot r_q(n)] + E[(r_q(n))^2]\}.$$

在大多数隐藏方案^[4,6,8,11,16,20,22]中, 秘密信息序列为 $w_k = 1$ 或 -1 , 对于 r_i^o 可以看成是 $I^o(n-1)$ 的前向预测系数 $I^o(n)$ 的预测误差, 所以 r_i^o 是一个零均值的高斯或 Laplacian 分布模型^[27], 即有 $E[r_i^o(n)] = 0$, 又因为原始信息序列与秘密信息序列是统计独立的. 因此, 我们得到 $E[r_i^o(n) \cdot r_q(n)] = E[r_i^o(n)] \cdot E[r_q(n)] = 0$, 那么有

$$\begin{aligned} E[GE_i^h] &= \frac{1}{N-1} \sum_n E[(r_i^o(n))^2] + \\ &\quad \frac{1}{N-1} \sum_n E[(r_q(n))^2] \\ &= E[GE_i^o] + \frac{1}{N-1} \sum_n E[(r_q(n))^2], \end{aligned}$$

因为 $E[(r_q(n))^2] \geq 0$, 因此有 $E[GE_i^h] \geq E[GE_i^o]$.

证毕.

对于二维离散序列的 $I(x, y)$, 定义水平梯度能量 GE_V 和垂直梯度能量 GE_H 如下:

$$GE_V = \frac{1}{N \times (M-1)} \sum_{x=1}^N \sum_{y=2}^M (I(x, y) - I(x, y-1))^2 \quad (6)$$

$$GE_H = \frac{1}{(N-1) \times M} \sum_{y=1}^M \sum_{x=2}^N (I(x, y) - I(x-1, y))^2 \quad (7)$$

总的梯度能量 GE 定义为

$$GE = GE_V + GE_H \quad (8)$$

同理, 掩密图像 $I^h(x, y)$ 的梯度能量均值 $E(GE_i^h)$ 大于等于干净图像 $I^o(x, y)$ 的梯度能量均值 $E(GE_i^o)$. 因此, 我们把图像的梯度能量作为空域图像的一特征.

3.1.2 梯度能量比特征

虽然从同一幅图像来看, 嵌入信息的掩密图像梯度能量比干净图像梯度大. 但是, 由于不同的图像梯度能量不同, 有图像本身梯度能量就很大, 而有图像梯度能量本身就很小的, 换句话说, 有的图像嵌入信息后的梯度能量远远小于另外原始图像的梯度能量, 所以用梯度能量作为一特征会把梯度能量大的干净图像误判为掩密图像, 把掩密图像梯度能量小的判为干净图像. 正因为此, 我们想到对图像的梯度能量进行归一化. 换句话说, 我们先对待测图像进行平滑攻击, 然后求攻击后的梯度能量, 与原来的梯度能量进行比值. 这样得到第二个特征, 即梯度能量比特征.

定理 2. 掩密信号序列 I^h 的梯度能量与其经

平滑攻击后的梯度能量比 RGE_i^h 大于等于原始信号 I^o 的梯度能量与其经平滑攻击后的梯度能量比 RGE_i^o , 且都大于等于 1, 即有 $RGE_i^h \geq RGE_i^o \geq 1$ 成立.

证明. 对于一维信号 I 不光滑的部分反映在频域里, 就是信号的高频部分. 而对于平滑攻击就相当于信号通过一个低通滤波器, 将滤去信号的高频部分. 也就是说信号的梯度能量经平滑攻击后将造成梯度能损失. 因此, 有 $GE_i \geq GE_{smooth}$ 成立 (GE_{smooth} 表示经平滑攻击后信号梯度能量), 梯度能量始终是一个大于零的数, 所以不管信号是否含有秘密信息都有 $RGE_i \geq 1$ 成立. 证毕.

梯度能量反映的是信号的光滑程度, 当梯度能量越大, 信号就越不光滑. 对于嵌入秘密信息, 常常可以看成是信号加入噪声, 如是 LSB 隐写, 相当于加入高斯白噪声, 而噪声的频率一般都比较低. 对于离散信号, 经平滑攻击或者说经低通滤波后, 高频部分将被滤掉, 而掩密信号的高频部分是大于原始信号的高频部分的. 因为原始信号嵌入秘密信息后, 其高频部分包括其自身高频和噪声频率. 所以掩密信号以平滑攻击后的梯度能量比原始信号经平滑攻击后梯度能量损失大, 它不光损失高频部分的梯度能量, 而且损失噪声的梯度能量. 换句话说, 掩密信号的经平滑攻击后的梯度能量近似地等于原始信号经平滑攻击后的梯度能量, 即 $E[GE_i^h] \cong E[GE_{smooth}^o]$ 成立 ($E[GE_{smooth}^h]$ 表示掩密信号经平滑攻击后的梯度能量均值, $E[GE_{smooth}^o]$ 表示原始信号经平滑攻击后的梯度能量均值). 又从上面知, 掩密信号的梯度能量的均值是大于等于原始信号的梯度能量的均值, 即有 $E[GE_i^h] \geq E[GE_i^o]$. 因此有

$$RGE_i^h = \frac{E[GE_i^h]}{E[GE_{smooth}^h]} \geq RGE_i^o = \frac{E[GE_i^o]}{E[GE_{smooth}^o]} \geq 1 \quad (9)$$

对于二维图像信号而言, 上式同样成立.

我们将图像的梯度能量特征与梯度能量比特征构成一个二维特征向量 (GE_i, RGE_i) .

3.2 DCT 域特征: DCT 系数能量和 DCT 系数能量差

3.2.1 DCT 系数能量特征

参照文献^[15-17, 19-20, 23, 26]和文献^[28], 基于 DCT 变换域的隐写方式通常可描述如下:

$$\begin{aligned} t^h(u, v) &= t^o(u, v) + J(u, v) \cdot \alpha \cdot w(u, v) \text{ 或} \\ t^h(u, v) &= t^o(u, v) \cdot (1 + J(u, v) \cdot \alpha \cdot w(u, v)) \end{aligned} \quad (10)$$

这里的 $t^o(u, v)$, $t^h(u, v)$ 分别表示原始图像和嵌入信息后图像的 DCT 系数, (u, v) 表示 DCT 系数的

位置, α 是控制因子, $w(u, v)$ 表示是秘密信息或水印信息(可能是二进制数也可能是实数), $J(u, v)$ 是强度阈值, 当其大于某个值时, 则肉眼能看出图像失真性^[20]. 如果式(10)左边式子的 $J(u, v) \cdot \alpha \cdot w(u, v)$ 的值为 1, 0, 或 -1 时, 则为 LSB 方式嵌入到 DCT 量化后的系数. 下面给出几个定义.

定义 1. 设图像的像素值为 $I(x, y)$, 则图像能量定义为

$$EN = \frac{1}{N \times M} \sum_x \sum_y E[(I(x, y))^2] \quad (11)$$

其中 N, M 表示图像的行数和列数, $E[\cdot]$ 表示均值.

定义 2. 设图像的 DCT 系数为 $t(u, v)$, 则图像的 DCT 系数能量定义为

$$DEN = \frac{1}{N \times M} \sum_u \sum_v E[(t(u, v))^2] \quad (12)$$

众所周知, 对于 JPEG 压缩过程中, DCT 变换本身不存在图像能量的损耗问题. 能量的损失主要出现在量化过程中, 而量化造成的图像能量损耗是无法补偿的. 对于 JPEG 格式的图像, 其解压后图像能量与 DCT 变换域中的能量是相同的, 即有 $EN = DEN$ 成立. 因此, 我们以后所讲的图像能量就是 DCT 变换域中的能量. 我们知道, 对于目前出现的 DCT 变换域隐写工具, 如 F5、Jsteg、Outguess、Jphide 等隐写工具都将秘密信息嵌入到量化后的 DCT 系数中. 秘密信息嵌入到量化后的 DCT 系数中, 必然会引起图像能量的改变.

定理 3. 直接嵌入图像 DCT 系数中的隐写过程会引起图像能量的增大, 即 $DEN_t^h > DEN_t^o$.

证明. 在式(10)的左边, 令 $\omega(u, v) = J(u, v) \cdot \alpha \cdot w(u, v)$, 则式(10)的左边等转变成

$$t^h(u, v) = t^o(u, v) + \omega(u, v) \quad (13)$$

原始图像能量为

$$DEN_t^o = \frac{1}{N \times M} \sum_u \sum_v E[(t^o(u, v))^2 \cdot (q(u, v))^2] \quad (14)$$

这里的 $q(u, v)$ 是在点 (u, v) 的量化步长, 是个常量.

掩密图像能量为

$$\begin{aligned} DEN_t^h &= \frac{1}{N \times M} \sum_u \sum_v E[(t^h(u, v))^2 \cdot (q(u, v))^2] \\ &= \frac{1}{N \times M} \sum_u \sum_v E[(t^o(u, v) + \omega(u, v))^2 \cdot (q(u, v))^2] \\ &= \frac{1}{N \times M} \sum_u \sum_v \{ (q(u, v))^2 \cdot E[(t^o(u, v))^2 + \\ &\quad 2 \cdot t^o(u, v) \cdot \omega(u, v) + (\omega(u, v))^2] \} \end{aligned} \quad (15)$$

我们知道, 秘密信息与图像信息是相互独立

的^[24], 所以有

$$E[t^o(u, v) \cdot \omega(u, v)] = E[t^o(u, v)] E[\omega(u, v)] \quad (16)$$

通常地, 对于嵌入的秘密信息一般都看成是一个具有零均值的加性白噪声, 也就是说 $E[\omega(u, v)] = 0$, 那么, 式(15)就转变为

$$\begin{aligned} DEN_t^h &= \frac{1}{N \times M} \sum_u \sum_v \{ (q(u, v))^2 \cdot \{ E[(t^o(u, v))^2] + \\ &\quad E[(\omega(u, v))^2] \} \} \\ &= DEN_t^o + \frac{1}{N \times M} \sum_u \sum_v \{ (q(u, v))^2 \cdot E[(\omega(u, v))^2] \} \end{aligned} \quad (17)$$

式(17)的最后部分 $(q(u, v))^2 \cdot E[(\omega(u, v))^2] > 0$, 所以 $DEN_t^h > DEN_t^o$, 定理 3 成立.

我们把图像能量, 也就是 DCT 系数能量当作 DCT 变换域的一个特征.

3.2.2 DCT 系数能量差特征

正如 3.1.2 节所讲的一样, 有的图像自身能量大, 而有的图像自身能量小, 对于能量小的图像, 即使再加上一定的噪声能量也只不过自身能量大的图像, 因此这种图像嵌入秘密信息后会很容易被误判为干净图像的. 反过来, 对于自身能量大的干净图像也会很容易被误判为掩密图像. 这样一来, 仅用一个特征来检测 JPEG 格式图像是否掩密是不够的.

我们想到用量化操作去噪^[29-30], 一个典型的信号(如图像)是结构相关的, 好的编码器利用结构相关性对数据进行压缩, 而且噪声没有结构冗余信息, 不容易被压缩. 因此, 一个好的数据压缩方法(量化方法)可提供一个适当的模型来识别信号和噪声. 对于量化去噪, 当系数的幅值小于量化步长时被置为零, 而大于门限的系数被进一步量化, 量化是数据压缩的关键步骤, 只要量化步长合适, 不会引起图像的显著失真. 量化也同时具有消噪功能. 基于以上的原理, 我们得出以下定理.

定理 4. 掩密图像 I^h 的 DCT 系数能量与其经量化攻击后的 DCT 系数能量差 $DDEN_t^h$ 大于等于原始信号 I^o 的 DCT 系数能量与其经量化攻击后的 DCT 系数能量差 $DDEN_t^o$, 即有 $DDEN_t^h \geq DDEN_t^o$ 成立.

证明. 由以上分析, 我们知道, 对量化操作具有消噪功能. 图像嵌入秘密信息过程相当于原始图像加入噪声过程. 当对掩密图像执行量化操作时, 其不但原有图像的噪声被消掉部分, 加入的秘密信息噪声也将被消去部分噪声, 所以其损失的能量显然比原始图像经量化攻击后损失的能量大. 也就有

$DDEN_I^h \geq DDEN_I^o$ 成立。证毕。

我们把图像的 DCT 系数能量特征和经量化攻击后的 DCT 系数能量差特征构成一个二维特征向量 $(DEN_I, DDEN_I)$ 。

4 分类器的设计

提取出的能反映具体问题的特征对于隐藏信息检测来说是一个至关重要的步骤,但没有一个好的、与其相适应的分类器,则分类效果也不会太好. 设计一个与特征相适应、与图像结构吻合的、具有最大识别率(也就是最小的虚警率和漏警率)的分类器也是一个至关重要的步骤. 对于空域或 DCT 变换域图像,当其特征被提取出来后,能否判定是否含有秘密信息是一个二类分类的问题. Manikopoulos 等^[23]

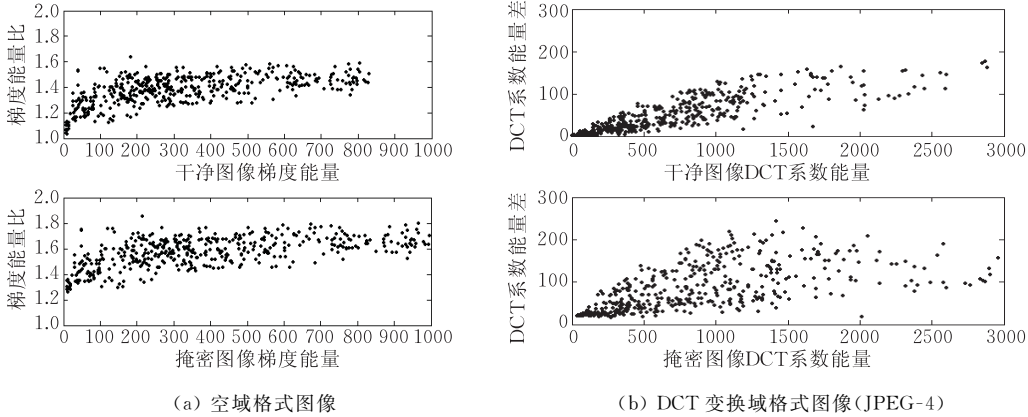


图 2 样本特征向量散列图像(注(a)和(b)的下图掩密图像都是从其上图干净图像嵌入秘密信息后所得的,所以其特征图分布形状相似)

在系统中我们采用了 Lei 等^[26]的三层前馈非线性神经网络来作为分类器. 在输入层是两个神经元,即是我们提取出来的二维特征向量(空域格式图像是 (GE_I, RGE_I) , DCT 变换域图像是 $(DEN_I, DDEN_I)$);输出层是一个神经元,即是分类结果有或无;中间层是由若干个神经元组成,用来记忆训练样本集的. 在多层神经网络中,我们采用 C 型螺线函数 $g(x)=1/(1+\exp(-x))$ 作为分类因子. 其值域为 $(0, 0, 1, 0)$, 当输出值靠近 1 时,就将测试图像判为掩密图像. 相反地,当输出值靠近 0 时,就将测试图像判为干净图像. 当然,分类前首先要决定是空域图像还是 DCT 变换域图像(JPEG 图像),因为不同格式的图像特征不同,分类器的隐层结构不同.

我们需要训练来设定分类器内部参数. 在训练阶段,每种样本(掩密图像特征向量样本和干净图像特征向量样本)训练误差经大量叠代后收敛到一个

提出使用二层神经网络去判别测试图像是否含有秘密信息. 当然,二层神经网络作为线性判别函数其效果是蛮好的. 但是,通常掩密图像与干净图像是非常相似的,如图 2 所示. 图 2(a)和(b)画出 500 幅干净图像和 500 幅掩密图像的特征. (a)图中上面是空域格式干净图像的二维特征,下面是掩密图像的二维特征. 从图像上可以看出,干净图像的特征点分布相对于掩密图像要集中些,干净图像梯度能量在 $0 \sim 850$, 梯度能量比在 $1 \sim 1.6$ 之间,而掩密图像梯度能量在 $0 \sim 1000$ 之间,梯度能量比在 $1.2 \sim 1.9$ 之间. (b)上图是干净 DCT 格式图像的二维特征点,下图是掩密图像的 DCT 格式图像的二维特征点,大部分干净图像的特征点,在 $[0, 2500, 0, 200]$ 范围,而掩密图像的特征点落在 $[30, 3000, 20, 250]$ 范围. 因此这种二类样本(掩密与干净)可采用二层神经网络进行分类.

定值. 当然,这并不意味着测试集的均方误差低于或接近训练集均方误差,就像过适应问题. 为了减小这个问题的影响,当训练到达一个最小误差时我们就停止训练.

对于彩色图像,则在每个空间都有一个二维特征向量,如空域格式图像,在 R、G、B 空间都有一个特征向量,在 DCT 变换域格式图像,在 Y、Cr、Cb 空间也都有一个特征向量. 对每个空间都要设定一个三层前馈非线性神经网络分类器. 根据选举法,如果二个空间内判为掩密或干净,则就判为图像掩密或干净.

5 仿真实验

5.1 实验准备

对于评估一个隐写分析算法的好坏,实验设计是很重要的. 实验设计的关键包括以下几点:

(1)通用性. 图像特征的提取和与之相应的分类器能够鉴别出测试图像是否含有秘密信息,不管这秘密信息是用哪种隐写工具嵌入的.

(2)性能好. 一方面,对掩密图像来说,具有高检测率;另一方面,对干净图像应保持较低的虚警率.

(3)鲁棒性. 分类器能够把经图像处理过的图像(如平滑、锐化、缩放、再压缩等)与掩密图像区分开.

基于以上考虑,我们选择 10 种隐写方式(嵌入到空域、DCT 变换域中的隐写算法各 5 种)作为我们实验评估. 空域隐写方式是 Ipv、Itp、s-tools、Forknox 和 Imagehide 5 种, DCT 变换域隐写方式是 Jphide、Jsteg、F5、Outguess 和 Jphswin 5 种.

我们知道,当一幅图像嵌入秘密信息量越多,就越容易检测出;反之,嵌入秘密信息越少则就与原始图像越难分开. 因此,嵌入率问题也是制约着正确分类率的一个重要因素. 所谓嵌入率(ER),就是嵌入的比特位与图像的像素量的比值,例如当秘密信息是 1k 字节嵌入空域图像时,如其像素量为 50k,则嵌入率 $ER = (1 \times 8) / 50 = 0.16\text{bpp}$ (bpp 是 bit per pixel 的缩写);当 8 比特位秘密信息嵌入到一个 8×8 大小的 DCT 块时,则嵌入率 $ER = 0.125\text{bpp}$. 我们用 10 种隐写方法嵌入的信息率不同,详细见表 1.

表 1 隐写工具与嵌入率

隐写工具	嵌入率/bpp	隐写工具	嵌入率/bpp
Ipv	0.33	Jphide	0.11
Itp	0.0156	Jsteg	0.02
s-tools	0.195	F5	0.13
Forknox	0.166	Outguess	0.08
Imagehide	0.21	Jphswin	0.3

本实验使用的图像库中有 4000 多幅原始彩色图像,其中 BMP、GIF 空域格式的图像各 1000 多幅, JPG 图像 2000 多幅. 用以上隐写工具嵌入信息后得到掩密图像各 1000 幅. 训练时采用空域图像 1000 幅干净图像(其中 BMP、GIF 各 500 幅), 2500 幅用以上空域隐写工具嵌入秘密所得的掩密图像;同样地, DCT 域图像 1000 幅干净图像和 2500 幅用以上 5 种 DCT 域隐写工具嵌入秘密信息后所得的掩密图像用作训练, 剩余图像用作测试图像, 并用 Photoshop 制作了空域图像和 DCT 域图像经平滑去噪、锐化、缩小、剪切(只做空域图像)、再压缩(只做 JPG 图像)各 100 幅,以供测试.

5.2 实验结果

本实验对空域图像、DCT 域图像进行不同方式提取特征,空域格式图像在 RGB 空间各提取两个特征,即每个空间里提取梯度能量和梯度能量比特特征构成二维特征向量;DCT 域格式图像在亮度空间 Y、红色差空间 Cr、兰色差空间 Cb 提取 DCT 系数能量和攻击后的能量比两个特征,构成一个二维特征向量,并把它们输入到神经网络分类器进行训练. 设计出不同格式、不同空间的三层非线性前馈式的神经网络分类器. 当对一幅图像进行测试判别时,其流程见图 3.

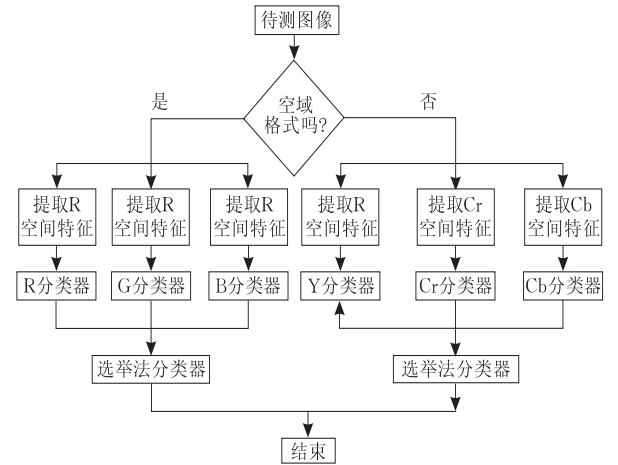


图 3 测试流程图

本实验对 7000 多幅图像(其中空域、DCT 域各 1000 幅, 10 种不同的隐写方式生成各 500 幅, 还有平滑、锐化、缩小、剪切(只做空域)、再压缩(只做 DCT 域)各 100 幅)进行测试, 测试结果见表 2. 表 2 中的操作域那一项中, 空域阳性检测率就是 5 种空域隐写方式的平均阳性检测率. 同样, DCT 域阳性检测率是 5 种 DCT 域隐写方式平均的阳性检测率. 下面介绍一些检测性能指标的概念.

阳性检测 PD(Positive Detection): 正确地识别掩密图像.

阴性检测 ND(Negative Detection): 正确地识别干净图像.

假阳性(或虚警)FP(False Positive): 将干净图像误判为掩密图像.

假阴性(或漏警)FN(False Negative): 将掩密图像误判为干净图像.

5.3 实验分析

从表 2 我们可以看出, 图像经处理后, 正确识别干净图像(ND)率明显降低, 而错误识别率则提高, 特别是图像经锐化攻击对 ND 率影响最大, 不管是

空域还是 DCT 域,ND 率都只有百分之六十几.相对而言,平滑攻击对图像的 ND 率影响最小.但都差于未经处理过干净图像的正确判别.

结合表 1 和表 2,我们发现,正确识别率不光与隐写方式有关,嵌入率也会对识别率有影响.空域中,Ipv 方式的嵌入率为 0.33bpp,其阳性检测率达到 96.80%,为所有隐写方式中的正确检测最高.DCT 域中,Jphswin 方式的嵌入率为 0.3bpp,嵌入率为 DCT 域中最高,其阳性检测(PD)率为 91.50%,为 DCT 域中的阳性检测率最高的.由此看来,嵌入率与阳性检测率是相关的.但我们也注意到,嵌入率与阳性检测率不是完全成线性关系的,如 s-tools 的嵌入率为 0.195bpp 高于 Forknox 的嵌入率 0.166bpp,但识别率却比 Forknox 的低.

表 2 不同隐写方式、不同操作域、不同处理和原始图像的平均 PD、ND、FN 和 FP 率

不同范畴		PD 率/%	FN 率/%
隐写方式	Ipv	96.80	3.20
	Itp	88.40	11.60
	s-tools	95.80	4.20
	Forknox	96.50	3.50
	Imagehide	95.60	4.40
	Jphide	76.20	23.80
	Jsteg	86.30	13.70
	F5	88.20	11.80
	Outguess	88.4	11.6
	Jphswin	91.50	8.50
操作域	空域平均	94.62	5.38
	DCT 域平均	86.12	13.88
不同范畴		ND 率/%	FP 率/%
原始或经处理过的干净图像类型	空域原始	87.30	13.70
	DCT 域原始	92.30	7.70
	空域平滑	81.0	19.0
	DCT 域平滑	85.0	15.0
	空域锐化	61.0	39.0
	DCT 域锐化	68.0	32.0
	空域缩小	68.0	32.0
	DCT 域缩小	77.0	23.0
	剪切(只含空域)	79.0	21.0
	再压缩(只含 DCT 域)	81.0	19.0

6 总结与展望

最近,信息隐藏技术已经在多个领域中应用,如版权保护、隐写术、指纹识别和数字版权管理等.虽然目前大多数研究都集中在怎样安全地、无视觉失真地嵌入信息.但是,对隐藏信息检测(即隐写分析)也已经风靡起来.在这篇论文里,提出了基于攻击的二维特征向量(空域是梯度能量和梯度能量比,DCT 域是图像 DCT 系数能量及其能量比)、神经网络分类器的隐藏信息盲检测方法.在我们的实验里,我们的图像数据库由几千幅原始干净图像和 10 种隐写工具嵌入的掩密图像组成,用这些图像来检验我们提取的特征和分类器的性能.

表 3 总结和比较了我们提出的特征方式同早期的文献著作的方式,我们可以归纳为以下几点:

(1)大多数系统提取的特征都来自空域、DFT 域和 DCT 域.

(2)我们与文献[11,25-26,31]相似的都是盲检测,也就是对隐写方式没有什么限制(如 LSB 或扩频等)

(3)对于难以分离的类集来说,采用非线性分类器比线性分类器更适合,我们系统、文献[23,26]都采用非线性的神经网络分类系统.

(4)从表 3 我们可以得出分类性能与所提取的特征数量没有必然的联系,但与隐写方式、嵌入率相关联.嵌入率更高更容易检测,嵌入率低更难检测;不同的隐写方式检测率也不一样.

(5)我们的训练与图像库收集了大量的干净图片和掩密图片作为样本,并且用不同的隐写方式(10 种)、不同嵌入率(0.02bpp~0.33bpp)和不同的图像处理(平滑、锐化、缩小、剪切和再压缩),这种多方位性使我们的系统更接近现实应用.

(6)我们提出的系统平均识别率(PD 率与 ND 率的总平均为 90.085%)比其他研究的识别率高.

表 3 早期著作与我们系统的总结比较

著作	特征数目	特征所在领域	分类器类型	隐写领域	测试方案数	掩密图像负荷/bpp	训练图数量	测试图数量	平均 PD 率/%	平均 ND 率/%
文献[11]	4	空域 DCT	线性	任意	6	>0.05	331	未报道	未报道	未报道
文献[13]	2	空域	线性	LSB	1	>0.05	未报道	80	97	未报道
文献[23]	64	DCT	神经网络	扩频	1	0.016	28	14	未报道	未报道
文献[25]	1	DFT	线性	任意	3	1	20	4	96.2	94.8
文献[26]	2	空域 DCT	神经网络	任意	6	0.01~2.66	1716	572	90.28	70.56
文献[31]	10	空域 DFT	线性	任意	6	>0.01	12	10	72.08	未报道
文献[32]	1	空域	线性	LSB	1	0.65	未报道	未报道	未报道	未报道
本文	4	空域 DCT	神经网络	任意	10	0.02~0.33	7000	7800	90.37	89.8

注:本文的平均 PD 率是空域 PD 率与 DCT 域 PD 率的总平均,即它们之和除以 2,本文的平均 ND 率计算与平均 PD 率一样计算.DFT 是指离散傅立叶变换.

为了使我们的系统更好、更实用,下一步的工作包括以下几点:

(1) 找到更有效的特征,能够很好地检测出绝大多数的隐写方式嵌入的掩密图像,对小波域图像特征做进一步的挖掘,以适应 JPEG2000 图像的盲检测.

(2) 使我们的算法能够适应不同的压缩算法图像(如 JPEG-x, H. 26x, MPEG-x).

(3) 能够鉴别各种隐写算法、估算出图像的掩密量、掩密区域,进一步地恢复出图像所含的秘密信息.

参 考 文 献

- [1] Jhnson N F, Jajodia S. Exploring steganography: Seeing the unseen. *IEEE Computer*, 1998, 31(2): 26-34
- [2] Zhang Li-He. Information detection and novel digital watermarking algorithms [Ph. D. dissertation]. Beijing: Beijing University of Posts and Telecommunications, 2004 (in Chinese)
(张立和. 隐藏信息检测与新型数字水印算法 [博士学位论文]. 北京: 北京邮电大学, 2004)
- [3] Katzenbeisser S, Petitcolas F A. *Information Hiding Techniques for Steganography and Digital Watermarking*. Norwood, MA: Artech House Press, 2000
- [4] Bender W, Gruhl D, Morimot N, Lu A. Techniques for data hiding. *IBM Systems Journal*, 1996, 35(34): 313-336
- [5] Nikolaidis N, Pitas I. Robust image watermarking in the spatial domain. *Signal Process*, 1998, 66(3): 385-403
- [6] Marvel L M, Boncelet C G J, Retter C T. Spread spectrum image steganography. *IEEE Transactions on Image Processing*, 1999, 8(8): 1075-1083
- [7] Lie Wen-Nung, Li Chun-Chang. Data hiding in images with adaptive numbers of least significant bits based on the human visual system//*Proceedings of the IEEE International Conference on Images Processing*. Kobe, Japan, 1999, 10: 286-290
- [8] Chen T-S, Chang C-C, Hwang M-S. A virtual image cryptosystem based upon vector quantization. *IEEE Transactions on Image Processing*, 1998, 7(10): 1485-1488
- [9] Petitcolas F A P, Anderson R J, Kuhn M G. Information hiding—A survey. *Proceedings of the IEEE, Special Issue on Protection of Multimedia Content*, 1999, 87(7): 1062-1078
- [10] Sayrol E, Vidal J, Cabanillas S et al. Optimum watermark detection in color images//*Proceedings of the 1999 International Conference on Image Processing*. Kobe, Japan, 1999, 2: 231-235
- [11] Fridrich J, Goljan M. Practical steganalysis of digital images—state of the art//*Proceedings of SPIE on Security and Watermarking of Multimedias Contents*. San Jos: Springer-Verlag, 2002, 4675: 1-13
- [12] Avcibas I, Memon N, Sankur B. Steganalysis using image quality metrics. *IEEE Transactions on Image Processing*, 2003, 12(2): 221-229
- [13] Kong X, Zhang T, You X, Yang D. A new steganalysis approach based on both complexity estimate and statistical filter//*Proceedings of the IEEE Pacific-Rim Conference on Multimedia*. New York: ACM Press, 2002, 2532: 434-441
- [14] Maes M. Twin peaks: The histogram attacks to fixed depth image watermark//*Proceedings of the Information Hiding*. London, UK: Springer-Verlag, 1998: 290-305
- [15] Huang J, Shi Y Q. Adaptive image watermarking scheme based on visual masking. *Electronics Letters*, 1998, 34(8): 148-150
- [16] Ogihara T, Nakamura D, Yokoya N. Data embedding into pictorial with less distortion using discrete cosine transform//*Proceedings of the ICPR'96*. Linz, Austria: IEEE, 1996, 2: 675-679
- [17] Cox I J, Kilian J, Leighton F T, Shamoon T. Secure spread spectrum watermarking for multimedia. *IEEE Transactions on Image Processing*, 1997, 6(12): 1673-1687
- [18] Podilchuk C I, Wenjun Z. Image-adaptive watermarking using visual models. *IEEE Journal on Selected Areas in Communications*, 1998, 16(4): 525-539
- [19] Wu M, Yu H, Lui B. Data hiding in image and video: Part II—Designs and applications. *IEEE Transactions on Image Processing*, 2003, 12(6): 696-705
- [20] Cheng Q, Huang T S. An additive approach to transform-domain information hiding and optimum detection structure. *IEEE Transactions on Multimedia*, 2001, 3(3): 273-284
- [21] Chen Y-S, Kwon O-H, Park R-S. Wavelet based watermarking method for digital images using the human visual system. *Electronic Letters*, 1999, 35(6): 466-478
- [22] Voloshynskiy S, Herrigel A, Rytsar Y, Pun T. Stego wall: Blind statistical detection of hidden data//*Proceedings of the SPIE*. California, USA: IEEE, 2002, 4675: 57-68
- [23] Manikopoulos C, Shi Y Q, Song S et al. Detection of block DCT-based steganography in gray-scale images. *Multimedia Signal Processing*, 2002, 12(1): 355-358
- [24] Chandramouli R. A mathematical approach to steganalysis//*Proceedings of the SPIE*. California, USA: IEEE, 2002, 4675: 14-25
- [25] Harmsen J J, Pearlman W A. Steganalysis of additive noise modelable information hiding//*Proceedings of the SPIE*. San-tanlara, California, USA: IEEE, 2003: 131-142
- [26] Lie Wen-Nung, Lin Guo-Shiang. A feature-based classification technique for blind image steganalysis. *IEEE Transactions on Multimedia*, 2005, 7(6): 1007-1020
- [27] Jain A K. *Fundamentals of Digital Image Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1989
- [28] Lie W-N, Lin G-S, Wu C-L. Robust image watermarking on the DCT domain//*Proceedings of the IEEE International Symposium on Circuits and Systems*. Geneva, Switzerland: IEEE, 2000: 1228-1231

- [29] Saito N. Simultaneous noise suppression and signal compression using a library of orthonormal bases and the MDL criterion//Foufoula-Georgiou E, Kumar P eds. Wavelets in Geophysics. New York: Academic, 1995: 299-324
- [30] Natarajan B. Filtering random noise from deterministic signals via data compression. IEEE Transactions on Signal Processing, 1995, 43(10): 2595-2605



MAO Jia-Fa, born in 1970, Ph. D. candidate, associate professor. His research interests include digital watermarking, steganalysis, pattern recognition, image and signal processing, and computer vision.

- [31] Avcibas I, Memon N, Sankur B. Steganalysis using image quality metrics. IEEE Transactions on Image Processing, 2003, 12(2): 221-229
- [32] Chandramouli R, Memon N. Analysis of LSB based image steganography techniques//Proceedings of the International Conference on Image Processing. Thessaloniki, Greece: IEEE, 2001: 1019-1022

LIN Jia-Jun, born in 1948, Ph. D. , professor and Ph.D. supervisor. His research interests include pattern recognition, intelligent control, and signal detection.

DAI Meng, born in 1979, Ph. D. candidate. His research interests include pattern recognition, and intelligent control.

Background

With the rapid development of digital multimedia and network technology, information hiding has been received much more attention both in theoretical and industrial fields during the last decades. The concept of information hiding is to hide data in the cover medium imperceptibly. Particularly, steganography is typical application of information hiding. The main purpose of steganography is for covert communication. In contrast to data hiding, steganalysis is the art of detecting the presence of hidden messages. It is developed to block the covert communication with illegal information for the urgent security demands of network. Steganalysis' technique is very important for the security of network information and how to search and detect secret messages which is transferred in network is crucial and practical to safeguard the national security. So Steganalysis' technique is an important part of the internet information security. For example, an inlet/outlet content-monitoring program can inspect and intercept suspected multimedia data which is transmitted on the network. In addition, steganalysis techniques can also be utilized to evaluate the security of the channels for covert communication under construction. The massive variety of data hiding methods make the design of staganlysis methods to cope with most embedding methods blindly and defiantly. Up to the present, hardly any detection method can explore

all kinds of embedding methods absolutely.

This work was supported by the Natural Science Foundation of Jiangxi Provincial Department of Education under the grant No. Gjj08462, and the Shanghai Education Development Foundation ("ChenGuang" project) under the grant No. 2008CGB21. The meaning of these projects is to research Steganalysis' technique, to develop and advance hidden messages blindly detect technique, to enhance the security of the network information for our country. This work was proposed that an attacked image was based on the hidden messages blindly detect technique. This technique is one of the most important parts of these project.

Since studying for Ph. D, the author has been done research which is in the area of hiding message blind detecting following his tutor. He and his research colleague have completed the Shanghai Security Research Institute 104 project of Image Hiding Message Detection System between November 2005 and October 2006. Since November 2006, they have taken part in Beijing Security Research Institute 115 project of research stego-image evaluating flat. So our work is supported by Beijing Security Research Institute 115 project. In last years, He and his research colleague have co-authored over 50 technical publications in these areas, including 8 journal papers and one patent.