中山大學 网络空间安全学院
SUN YAT-SEN UNIVERSITY SCHOOL OF CYBER SCIENCE AND TECHNOLOGY

多媒体安全实验报告

实验内容：隐写算法实现与卡方检测

实验时间：2024/3/20

姓　　名：林恒盛

学　　号：21312246

目录

一 实验目的 **3**

二 实验任务 **3**

三 实验原理 **3**

1. LSB 隐写算法 3
2. 卡方分布 4
3. 卡方检测 5
4. LSB 图像的卡方检验 6

四 实验分析 **6**

1. LSB 隐写算法的基本实现 6
2. 用卡方分布检测进行隐写分析 6
3. LSB 的改进 8

五 实验总结 **9**

隐写算法实现与卡方检测

一　实验目的

1. 介绍并实现算机领域中最低有效位（LSB）实现的基本概念和技术
2. 理解 LSB 实现的基本原理和技术
3. 学习如何利用 LSB 实现信息隐藏，并利用卡方分布检测的方法进行隐写分析

二　实验任务

1. LSB 隐写算法的基本实现
2. 用卡方分布检测进行隐写分析
3. LSB 的改进

三　实验原理

1.　LSB 隐写算法

LSB 全称为 Least Significant Bit（最低有效位），是一种常被用做图片隐写的算法。

LSB 算法在空域上实现信息的嵌入，即将信息转换而成的二进制比特串嵌入到图像中像素位的最低位。在 LSB 域上的修改对于人眼是不易察觉的，因此具有较好的视觉效果。但由于 LSB 域同样在各种操作上被广泛利用，如压缩编码等，这导致 LSB 隐写算法的鲁棒性较差，容易导致嵌入信息被毁坏。

对于一张普通的 RGB 图像，它的像素值可以表示成这样的形状：$(height, width, channel)$(Figure 1).

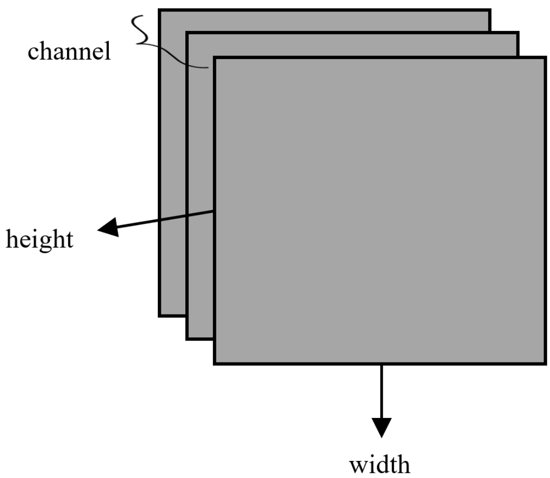

Figure 1: General Shape of RGB Image's Pixel

不同模式的图像仅仅是 channel 在数量上的区别。因此，对于图像的每个通道，都可以进行如 Figure 2 的空域表示，而 LSB 隐写算法便是在最底层平面 (LSB space) 上进行信息的嵌入。

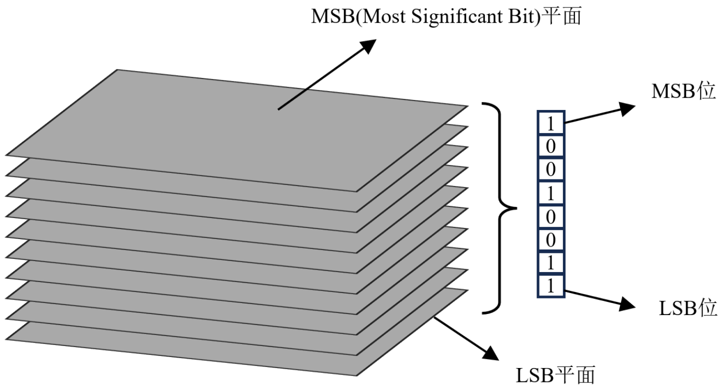

Figure 2: Bits Space of a Channel

2. 卡方分布

若 n 个相互独立的随机变量 $Z_1, Z_2, \cdots, Z_n$ 服从正态分布，则它们的平方和 $Z_1^2 + Z_2^2 + \cdots + Z_n^2$ 服从自由度为 n 的卡方分布：

$$X = \sum_{i=1}^{n} Z_i^2, \qquad X \sim \chi_n^2$$

对于一张图像的像素值，显然，它要么是偶数，要么是奇数，对于 0-255 的每个像素值，它在 LSB 隐写中服从下列变换：

$$F_{LSB} = 0 \leftrightarrow 1, 2 \leftrightarrow 3, \cdots, 254 \leftrightarrow 255$$

对于上述 LSB 变换的每个像素对 P，服从伯努利分布，即经过 LSB 变换，得到的像素对要么是像素对中的奇数，要么是像素对中的偶数，相应地，LSB 位要么为 0，要么为 1. 不妨设 LSB 为 1 的概率是 p.

根据中心极限定理，n 重伯努利分布中，试验成功的次数在 n 趋向于无穷时呈正态分布。设 $X = k$，即试验 n 次，恰好有 k 次成功的随机变量，有：

$$\lim_{n\to\infty} P\left(\frac{X - np}{\sqrt{np(1-p)}} \le x\right) = \Phi(x)$$

由上述公式可以得到，在 n 足够大时，存在：

$$\lim_{n\to\infty} P(x_i \ge X \le x_j) \approx \Phi(y_j) - \Phi(y_i),$$

$$\text{其中 } y_j = \frac{x_j - np}{\sqrt{np(1-p)}}, y_i = \frac{x_i - np}{\sqrt{np(1-p)}}$$

一张图像中，每个像素对的重复数量（即伯努利试验的重复次数）是足够的：一般规定，$p < q, np \ge 5$ 或者 $p > q, np \ge 5$ 时，多重伯努利分布即可使用正态分布近似进行概率计算。

因此每个像素对中偶数或奇数的个数近似服从正态分布。设每个像素值的频数为：$H(x)$，则 H_{2i} 或 H_{2i+1} 在 LSB 变换过程中近似服从正态分布。

对于每个像素，它为偶数或者为奇数的概率是一定的，即单重伯努利分布的概率 p 是确定的，不妨设 $p = \frac{1}{2}$, 则对每个像素对，随机变量 X 与常量 H_{2i} 为该像素对中像素为偶数的个数，则 n 为 $H_{2i} + H_{2i+1}$, 令 $H_{2i}^* = H_{2i} + H_{2i+1}$, 有：

$$\begin{aligned}
P(X \geq H_{2i}) &\approx 1-\Phi(y_{2i}) \\
&= 1-\Phi(\frac{H_{2i}-np}{\sqrt{np(1-p)}}) \\
&= 1-\Phi(\frac{H_{2i}-H_{2i}^{*}p}{\sqrt{H_{2i}^{*}p(1-p)}}) \\
&= 1-\Phi(\frac{H_{2i}-\frac{1}{2}H_{2i}^{*}}{\sqrt{\frac{1}{4}H_{2i}^{*}}}) \\
&= 1-\Phi(\frac{2H_{2i}-H_{2i}^{*}}{\sqrt{H_{2i}^{*}}})
\end{aligned}$$

由上述公式可以得到，变量 $r_i = \frac{2H_{2i}-H_{2i}^{*}}{\sqrt{H_{2i}^{*}}}$ 服从正态分布。因此随机变量 $z=\sum_{i=0}^{127} r_i^2$ 服从自由度为 128 的卡方分布。

对于自由度为 k 的卡方分布，其概率密度函数如下（$x \leq 0$ 时，$f(x;k)=0$）：

$$f(x;k)=\frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})}x^{\frac{k}{2}-1}e^{-\frac{x}{2}}$$

代入公式，可以得到 LSB 替换中，随机变量 z 的概率密度函数如下：

$$f(z;k)=\frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})}z^{\frac{k}{2}-1}e^{-\frac{z}{2}}\ ,\ z=\sum_{i=0}^{127}(\frac{2H_{2i}-H_{2i}^{*}}{\sqrt{H_{2i}^{*}}})^2$$

$$\Gamma(x)=\int_0^{\infty} t^{x-1}e^{-t}dt$$

变量 $z=\sum_{i=0}^{127}(\frac{2H_{2i}-H_{2i}^{*}}{\sqrt{H_{2i}^{*}}})^2=\sum_{i=0}^{127}(\frac{2H_{2i}-(H_{2i}+H_{2i+1})}{\sqrt{H_{2i}^{*}}})^2=\sum_{i=0}^{127}(\frac{H_{2i}-H_{2i+1}}{\sqrt{H_{2i}^{*}}})^2$. 由于图像经过 LSB 替换之后，会得到不变的 $H_{2i}^{*}=H_{2i}+H_{2i+1}$ 以及下降的 $|H_{2i}-H_{2i+1}|$, 因此随机变量 z 越小，表示存在 LSB 隐写的可能性越大。

卡方分布的累积分布积分函数 CDF 如下：

$$\begin{aligned}
& P(z \leq Z)=\int_{-\infty}^{Z} f(z;k)dz \\
\Rightarrow\ & P(z \leq Z)=\int_{0}^{Z} f(z;k)dz \\
\Rightarrow\ & P(z \leq Z)=\frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})}\int_0^Z z^{\frac{k}{2}-1}e^{-\frac{z}{2}}dz \\
\Rightarrow\ & P(z \leq Z)=\frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})}\int_0^Z 2^{\frac{k}{2}-1}(\frac{z}{2})^{\frac{k}{2}-1}e^{-\frac{z}{2}}dz \\
\Rightarrow\ & P(z \leq Z)=\frac{1}{\Gamma(\frac{k}{2})}\int_0^Z (\frac{z}{2})^{\frac{k}{2}-1}e^{-\frac{z}{2}}d\frac{z}{2} \\
\Rightarrow\ & F_k(x)=\frac{1}{\Gamma(\frac{k}{2})}\gamma(\frac{k}{2},\frac{x}{2})
\end{aligned}$$

上述公式中，γ 表示下不完全的 Γ 函数。由于随机变量 z 越小，表示存在 LSB 隐写的可能性越大。因此，$1-F_k(z)$ 可以衡量出现 LSB 隐写的可能性。

3. 卡方检测

卡方检测是一种假设检验方法，用于确定观察到的数据与期望数据之间的差异是否显著。由于卡方检测的目的是为了衡量观测值与期望值的误差，因此卡方分布的自由度往往比真实分布小 1. 此外，卡方检测分为两种：卡方拟合优度检验和卡方独立性检验。

A. 卡方拟合优度检验用于检验一个样本数据集是否符合一个理论分布，如正态分布、均匀分布等。

B. 卡方独立性检验则是为了确定两个变量之间是否存在关联。它通过计算观察数据的期望频数，再将期望频数与观察频数进行比较完成。

4. LSB 图像的卡方检验

由于图像的像素对服从卡方分布，而 LSB 隐写算法导致随机变量 z 降低，因此 z 越小，越有可能存在 LSB 隐写。可以采用卡方拟合优度检验对 LSB 隐写算法进行分析。

此处的卡方拟合优度检验实际上是计算 $F_k(z)$ 的值。当该值小于显著性阈值 0.05 时，说明随机变量 z 处于该当量的可能性是极低的。但此时图像却表现出了这种极低的可能，说明图像极有可能存在 LSB 隐写。

由于卡方拟合优度检验返回的是差异，因此会得到 $p = 1 - F_k(z)$，此时，得到的 p 越大，图像越有可能存在 LSB 隐写。

四 实验分析

1. LSB 隐写算法的基本实现

LSB 连续隐写替换可以通过先将图像中的 LSB 位置 0，然后再将二进制表示的嵌入数据写入 LSB 位。在提取嵌入信息的时候，往往需要确定结束位置，因此可以自定义结束标志，一同嵌入 LSB 平面中。

LSB 隐写信息的提取中，通过将 LSB 平面的二进制数据重组，再根据快速的字符串匹配算法 KMP 进行字符串匹配，找到既定的 flag，将 flag 前的二进制数据转化回原始数据形式即可。

Figure 3 是 lenna.jpg 图像嵌入随机比特串前后的像素分布直方图。所嵌入的数据由 numpy 随机生成，覆盖整个 LSB 平面。可以看到，使用随机比特串替换 LSB 平面后，图像出现了一定程度上的“削峰”现象，即值对效应。但由于原始图像为 jpeg 图像，经过一定的压缩编码，像素分布直方图的峰值并不算突出。尽管如此，LSB 隐写造成的值对效应依旧明显。

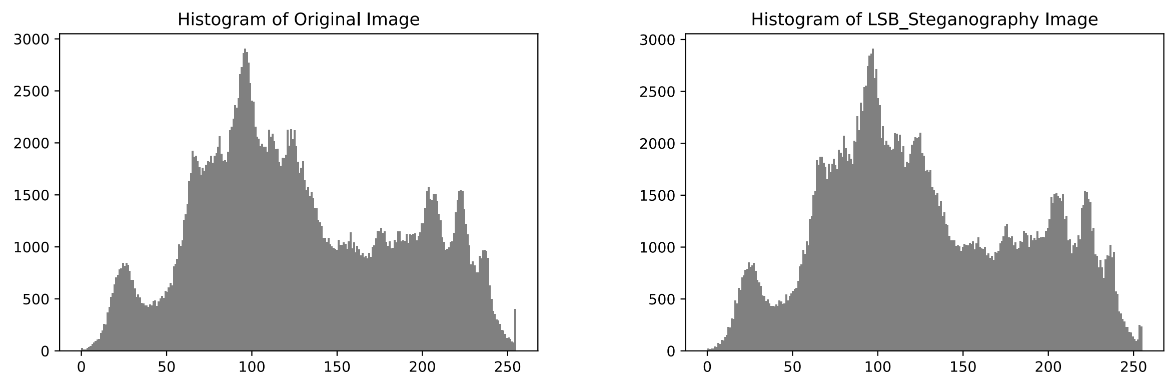

Figure 3: Histogram of Pixels before and after LSB

2. 用卡方分布检测进行隐写分析

此处使用 numpy 生成 100 个 0 到 1 均匀分布的数据，作为 LSB 隐写的嵌入率，即所嵌入随机比特串占 LSB 平面总像素个数的比例。

对不同的嵌入率进行 LSB 连续替换，然后使用卡方检测计算每张隐写图像的卡方统计量与隐写概率 p 值如 Figure 4 所示。

Figure 4 表明，在嵌入率达到 50% 左右，值对效应开始明显起来；在嵌入率达到 60% 或以上时，值对效应已经非常明显。并且此时，卡方检测的结果是显著的。这也说明了 LSB 连续替换易被检测的缺点。

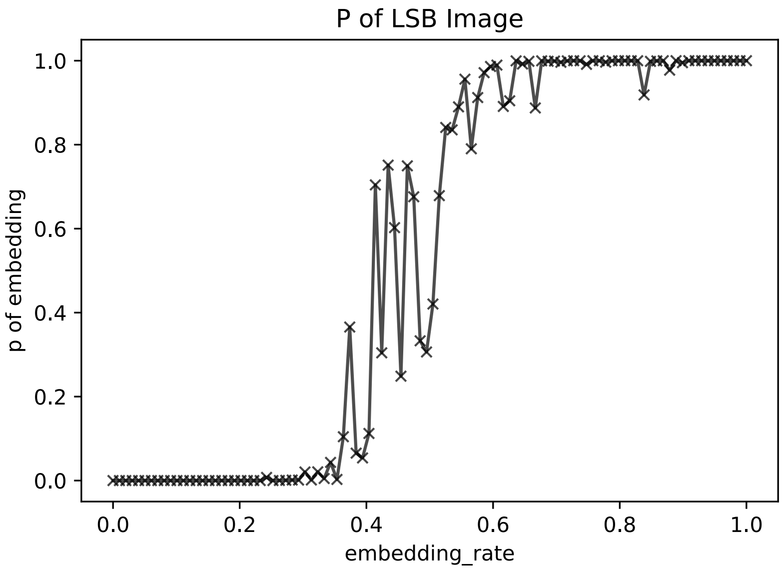

Figure 4: Probability of LSB Image in Different Embedding Rate

对图像像素的平均奇偶差 $|H_{2i} - H_{2i+1}|$ 进行统计，得到 Figure 5. 可以看到，嵌入率越高时，平均奇偶差越低，这也说明了理论的推测是正确的。

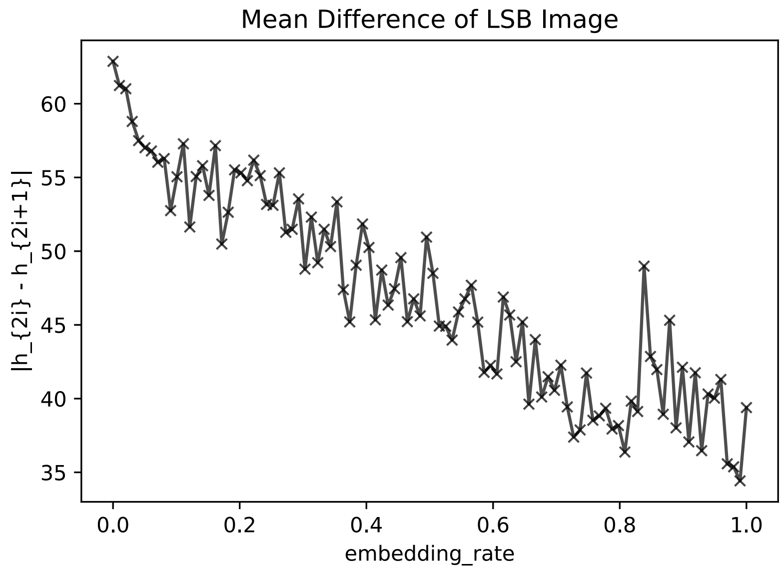

Figure 5: Average Parity Difference of Different Embedding Rate

3. LSB 的改进

从上述卡方检测可以看出，连续 LSB 隐写替换造成的值对效应是显著的。因此，此处使用随机位置的 LSB 替换进行实验。

随机位置的 LSB 替换通过 numpy 的固定随机种子进行，先生成 LSB 平面范围的随机位置索引，再将隐写信息按照随机索引进行嵌入即可。通信双方需要事先约定好随机种子，以使嵌入与提取时产生的随机索引是一定的。

随机嵌入得到的图像经过卡方检测后，统计 p 值如 Figure 6 所示. 对比连续 LSB 隐写替换，随机位置的 LSB 替换可以明显减轻低嵌入率 ($\leq 50\%$) 下 LSB 隐写的值对效应。但仍旧无法避免高嵌入率带来的显著性。

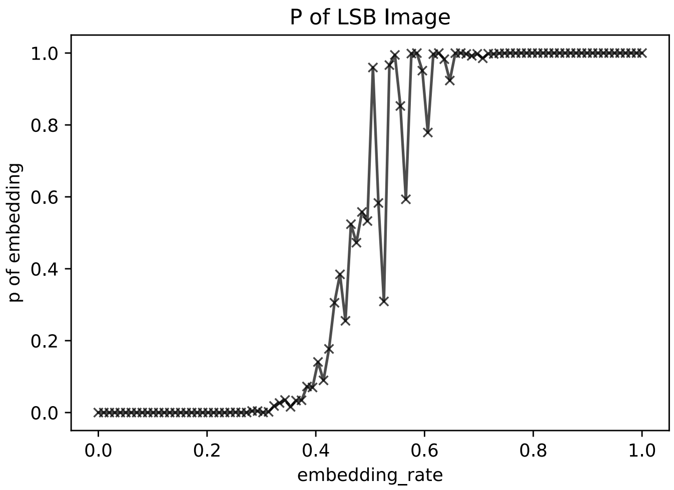

Figure 6: Probability of Random LSB Image in Different Embedding Rate
$randomseed = 1024$

对图像像素的平均奇偶差 $|H_{2i} - H_{2i+1}|$ 进行可视化如 Figure 7 所示。同样可以看到，随机位置替换在一定程度上减缓了平均奇偶差的下降过程。

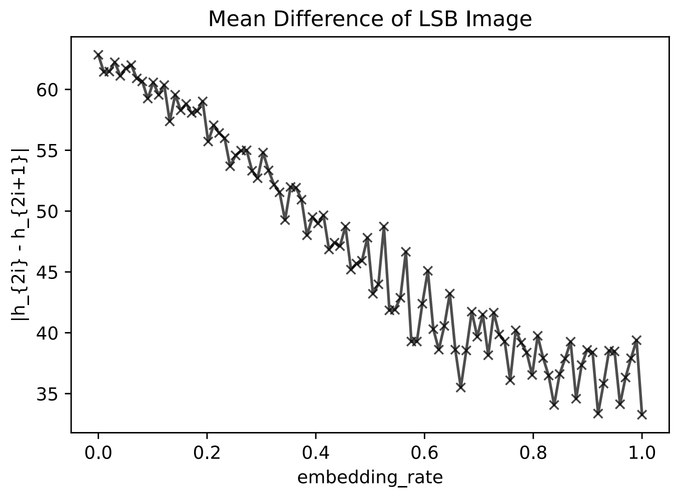

Figure 7: Average Parity Difference of Different Embedding Rate
$randomseed = 1024$

五 实验总结

本次实验进行了 LSB 连续隐写替换与随机隐写替换的 python 实现，并成功应用了卡方检测进行隐写分析，加深了对 LSB 隐写技术的认知和基本的统计检测原理与方法应用。

实验结果表明，无论是连续隐写替换还是随机隐写替换，嵌入率达到 50% 以上都不是一个良好的选择，容易被卡方检验分析出是否进行了隐写。这也同样说明，LSB 隐写替换作为实际的隐写应用时，嵌入效率并不理想，即可用于装载信息的像素位占比较小。

进一步改进 LSB 隐写替换的可能方法有在多个轻显著性位平面上进行随机隐写替换、增加跳变范围等。这既涉及到位选择的随机性，又囊括了嵌入位置的随机性，理论上有助于减弱值对效应的影响。