数字水印中的攻防分析*

孙水发[1], 仇佩亮[1], 张华熊[2]
(1. 浙江大学信息科学与电子工程学系, 浙江 杭州 310027;
2. 浙江理工大学信息电子学院, 浙江 杭州 310033)

摘 要: 从数字水印需要传递水印及原始媒体两个信息出发研究了数字水印中各个角色的攻防关系。利用了数字水印的不可感知性这一要求, 提出了图像空间, 感知空间及检测空间的概念, 并在这 3 个空间上定义了图像空间阈值, 感知空间阈值和检测空间阈值 3 个阈值。分析了数字水印系统中发送者、攻击者及接收者在上述 3 个阈值之间的攻防关系, 结果表明, 数字水印与大多数信息隐藏技术一样, 其实是一个寻找隐蔽信道的问题。在此基础上给出了具体例证, 得出了一些有益于数字水印嵌入、检测及攻击的指导性思想。

关键词: 数字水印; 隐蔽通信; 隐蔽信道; 信息论模型
中图分类号: TP391 文献标识码: A 文章编号: 0529-6579 (2004) S2-0094-04

自从现代意义的数字水印文章发表以来[1], 对数字水印的研究逐步扩展和深入[2-6]。然而我们发现很多水印系统只考虑传递水印信息, 而忽略了传递原始媒体数据的重要性。最初的水印系统将原始的媒体数据当作噪声: 在非盲的水印系统中[1, 2], 接收方通过将接收到的信息与原始的媒体数据相减从而得到水印信息; 在盲水印系统中[3], 接收方通常根据媒体数据噪声模型设计相应的最佳接收机实现检测。Cox 等[5] 从水印的检测方出发, 将宿主媒体看成通信系统的状态信息, 据此得出数字水印是一个带有边信息的通信过程的结论。Moulin[6] 则从香农的信息理论出发, 得出了最佳攻击策略是解决一个特定的率失真问题, 而最佳嵌入策略则是解决一个信道编码问题的结论。

以上模型及理论, 都忽略了传递原始媒体数据的重要性, 对水印的不可感知性要求重视不够。本文利用了数字水印的不可感知性这一要求分析了数字水印的通信系统模型, 研究了数字水印中各个角色的攻防关系, 得出了一些有益于数字水印嵌入、检测及攻击的指导性思想。

1 数字水印的通信系统模型

一个完整的数字图像水印的通信系统模型如图 1 所示, 涉及 3 个角色, 这里我们依次假设为发送者 Alice, 接收者 Bob 和攻击者 Eve。三者的任务分别是:

①Alice 希望嵌入信息后得到的图像 c_W 与原始图像 c_0 的差别应该是不可感知的, 而且 Bob 能从中恢复出原始消息 m。

②Eve 则总是希望对图像 c_W 做某些改动后, 既能使得到的图像 c_W' 与 c_W 的差别是不可感知的, 又使接收者 Bob 不能从中恢复出原始消息 m。

③Bob 应该从一幅嵌入了水印的图像 c_W 或 c_W' 中恢复出原始消息。

为了方便后面的讨论, 接下来先给出一些相关术语与定义。

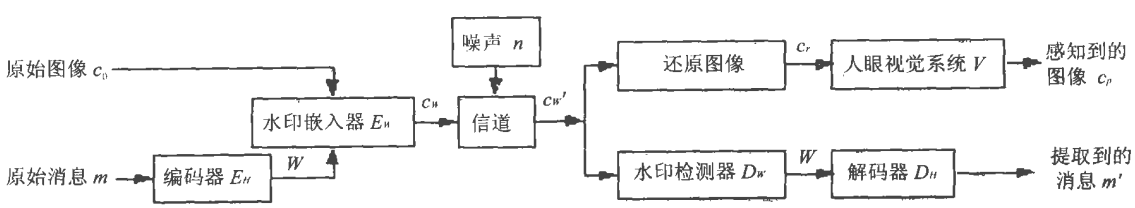

图 1 完整的数字图像水印的通信系统模型

Fig. 1 A complete communication model of digital image watermarking

* 收稿日期: 2004-09-11
作者简介: 孙水发 (1977 年生), 男, 博士研究生; E-mail: watersun1977@163.com

2　术语及定义

2.1　3 个空间

(1) 一个 $p\times q$ 的像素矩阵确定了一个维数 $M=p\times q$ 的图像空间 $\mathcal{I}$。

(2) 感知空间 $\mathcal{C}$ 是在图像空间 $\mathcal{I}$ 上的一个一对一的映射，一种感知空间可以认为是图像通过人眼的视觉系统后得到的感觉。

(3) 而长度为 L 的水印序列 W 来自维数为 L 的水印空间 $\mathcal{L}$。

2.2　3 个距离

接下来定义在这 3 个空间上的 3 个距离。

(1) D_H^{c-d} 为图像空间 $\mathcal{I}$ 上 2 个点 c、d 之间的距离。

(2) D_V^{i-j} 为感知空间 $\mathcal{C}$ 上 2 个点 i、j 之间的距离。

(3) D_W^{x-y} 为水印空间 $\mathcal{L}$ 上 2 个点 x、y 之间的距离。

2.3　3 个阈值和 3 个集合

有了以上 3 个空间及 3 个距离的概念，再定义 3 个阈值 $T_V^{i_0}$、$T_H^{c_0}$、$T_W^{x_0}$ 以及由这 3 个阈值确定的 3 个集合 H_{i_0}、H_{c_0}、H_{x_0}。

2.3.1　感知阈值 $T_V^{i_0}$ 与感知集合 V_{i_0}　$T_V^{i_0}$ 表示在感知空间 $\mathcal{C}$ 上的点 j 与点 i_0 距离 $D_V^{j-i_0}$ 大于 $T_V^{i_0}$ 时，则可以认为 i_0 与点 j 所对应的图像空间的图像来自不同的 2 幅图像，否则，可以认为点 i_0 与点 j 所对应的图像空间的图像来自同一幅图像。在感知空间 $\mathcal{C}$ 上所有与点 i_0 距离 $D_V^{j-i_0}$ 小于 $T_V^{i_0}$ 的点组成的集合用 V_{i_0} 表示。

2.3.2　图像集合 H_{i_0}、图像阈值 $T_H^{c_0}$ 和图像集合 H_{c_0}　感知空间的点集 V_{i_0} 映射到图像空间，可以得到相应的图像空间集合 H_{i_0}，则集合中在图像空间中距离最大的 2 点之间的距离称为图像阈值 $T_H^{c_0}$，其中 i_0 为 c_0 在感知空间上的映射。而由图像空间中所有与 c_0 的距离都小于 $T_H^{c_0}$ 的点组成的集合用 H_{c_0} 表示。

2.3.3　水印空间阈值 $T_W^{x_0}$，水印空间集合 W_{x_0} 和图像空间集合 H_{x_0}　$T_W^{x_0}$ 表示在水印空间 $\mathcal{L}$ 上的点 y 与点 x_0 距离 $D_W^{y-x_0}$ 大于 $T_W^{x_0}$ 时，则可以认为 x_0 与 y 为 2 个不同的水印，否则，认为点 y 与 x_0 为同一个水印。在水印空间 $\mathcal{L}$ 上所有与点 x_0 距离 $D_W^{y-x_0}$ 小于 $T_W^{x_0}$ 小于的点组成的集合用 W_{x_0} 表示。集合 W_{x_0} 映射到图像空间时，取能从中提取与相应水印 x_0 无法区分的图像点集，用 H_{x_0} 表示。

3　水印的攻防游戏

根据第 1 节的数字图像水印的通信系统模型以及第 2 节的相关定义，假设原始图像为 c_0，i_0 为 c_0 在感知空间上的映射，给定感知阈值 $T_V^{i_0}$，则可以得到图像空间 $\mathcal{I}$ 上与映射到 i_0 的图像 c_0 认为来自同一幅图像的集合 H_{i_0}，进而可以得到阈值 $T_H^{c_0}$ 和图像集 H_{c_0}。而对于水印 x_0，给定水印阈值 $T_W^{x_0}$，则可以得到图像空间 $\mathcal{I}$ 上与在图像空间上能从中提取与水印 x_0 认为相同水印的图像的集合 H_{x_0}。这 3 个图像集 H_{i_0}、H_{c_0} 和 H_{x_0} 的关系如图 2 所示。发送方 Alice 要做的事情是如何实现她的水印嵌入器 E_W，使得嵌入水印后的图像 c_W 处于集合 H_{c_0}、H_{i_0} 和 H_{x_0} 的交集 7 区，并且尽可能的增大 $T_H^{c_0}$ 的值，以期嵌入更多的消息；攻击方 Eve 总希望她将噪声 n 加入 C_w 后得到的 C_w' 处于 4 区或者 3 区。如果 Alice 增加了 $T_H^{c_0}$ 的值，以便嵌入更多的信息，则必将增加 H_{i_0} 区的范围，从而 4 区和 3 区也将相应增大，从而 Eve 有更大的攻击范围。检测方 Bob 则努力实现他的水印检测器 D_W，增加 H_{x_0} 的范围，一方面增加了 7 区的范围，从而为 Alice 嵌入信息提供更多的可选范围，另一方面也将减少 4 区和 3 区的范围，减小 Eve 的攻击范围。

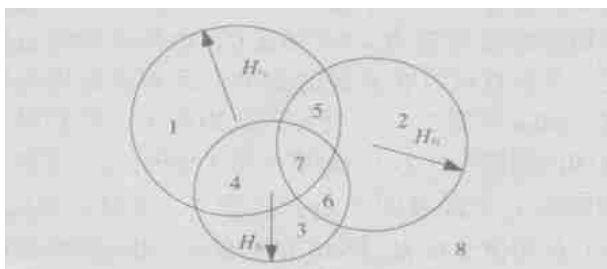

图 2　图像集合 H_{c_0}、H_{i_0} 和 H_{x_0} 的关系

Fig. 2　The relationship of image sets H_{c_0}、H_{i_0} and H_{x_0}

4　数字水印的攻防例证

正如我们前面指出，目前的很多水印系统模型忽略了传递另一个消息——原始图像 C_0 的重要性。但从图 1 可看出，在实际的水印通信系统中，发送方要传递的信息不只是原始水印消息 m，还包括原始的图像数据 c_0，在接收端不只要检测出消息 m，而且要给出一个还原的图像数据 c_r。而且从第 3 节的分析可知，不可感知性的水印的攻防

游戏中扮演着重要角色。

4.1 数字水印与隐蔽信道通信

从 Herodotus 描述的公元前 440 年，Histiaus 通过将信息写到奴隶的光头上，再让头发长起来实现隐蔽通信[4]，到现代 Tirkel 等[1]提出的基于 LSB（least significant bit，最不重要位）的水印嵌入技术，以及 Cox 等[2]提出的基于扩频的数字水印技术，其实都是在寻找一个隐蔽的信道，使得在前面描述的攻防游戏中攻击者无法实现有效的攻击，从而达到隐蔽通信的目的。

为了说明数字水印与传统的隐蔽通信（比如前面 Histiaus 的通信方式）之间的关系，我们给出一个图像水印的例子。图 3（a）是标准的 8 位 256×256 的 Lena 灰度图像，对应于图 1 中的原始图像 c_0，其 LSB 为第 8 位。图 3（b）是图 3（a）的 LSB 显示，其中黑点表示该位为 0，白点表示该位为 1。图 3（c）是图 3（a）去除其 LSB 后的图像。对于给定的 T^{i}_{Ψ} 和 T^{x}_{Ψ}，则根据第 2 节的定义，可以得到 3 个图像区域，H_{i_0}，H_{c_0} 及 H_{x_0}。这里假设图 3（c）处于 H_{c_0} 及 H_{i_0} 的交集内，则如果 Alice 将图 3（c）传给 Bob，Bob 将无法感知到其与图 3（a）的差别。这样，Alice 可以利用空余出来的 LSB 信道传递她所想传递的任何信息。也就是说，Alice 从原始的宿主图像 c_0 中构造出一个容量 $C=65536$ 的隐蔽信道。Alice 选取一个符合独立，均匀分布的随机矩阵 x_0，如图 3（d）所示，将此矩阵作为要传输的水印，对应于图 1 中的 W，而且可以知道该矩阵的信息量 $H=65536\leqslant C$，根据香农的信息论，该信息可以在上述的信道上无差错的进行传送。Alice 将图 3（c）叠加到此矩阵上，得到嵌入水印的图像图 3（e），对应于图 1 中的 c_W。现在水印矩阵 x_0 已经被图 3（c）隐藏了，也就是说图 3（e）的图像也在 H_{c_0} 及 H_{i_0} 的交集内。Bob 接收到图 3（e）的图像后，可以直接提取其 LSB 作为提取到的水印 W'。显然，此时提取到的所有位都是正确的，因此图 3（e）还处于内 H_{x_0}，这说明此时图 2 的 7 区存在。Alice 成功地实现了她的水印嵌入器 E_W，Bob 的水印检测器 D_W 也不错，可以正确的检测出水印。这跟 Histiaus[4] 的做法其实是一样的，先找到一个隐蔽的信道，将信息写入，然后加上一些伪装，发送出去。

4.2 隐蔽信道的攻防

上面的水印算法所建立的隐蔽信道很容易受到攻击。假设 Eve 也得到了一个图 3（e）的拷贝，她如何进行有效的攻击，使得检测方 Bob 无法检测到 Alice 嵌入的水印呢？

图 3 LSB 的水印嵌入、攻击及检测

Fig. 3 Digital watermarking algorithm in the LSB of an image: embedding, attack and detection

情况 1：Eve 对 Alice 的嵌入方法一无所知，无法找到那个隐蔽的信道。

Eve 尝试用 JPEG 压缩方法，品质因数为 95，得到如图 3（f）所示的压缩后的图像，对应于图 1 中的 c_W'。现在假设 Bob 无法区分 c_W' 与 c_W，如果 Bob 仍然用他刚才的检测器——提取图像 c_W' 的 LSB 作为检测到的水印，则他将发现 Alice 嵌入的 65 536 bit 的水印序列有 326 43 bit 错误，误码率为 49.8%，可以说整个信道差不多都被 Eve 成功破坏了，不能传输任何信息，也即此时 c_W' 处于图 2 的 3 区或 4 区。但此时，Eve 要冒一个险，她可能将 c_W' 推出 H_{i_0} 区，也即 Bob 可以区分 c_W 与 c_W' 了，从而攻击失败。

情况 2：Eve 知道 Alice 的嵌入方法，可以准确的找到那个隐蔽的信道。

此时，Eve 可以对该信道进行任意的破坏，而且不用冒着 c_W' 将推出 H_{i_0} 区的危险。

在以上 2 种情况下，Eve 都可能成功实现攻击。但 Bob 亦可以阻止 Eve 的攻击。假如他的检测器能够在传输的 65 536 bit 中有 32 643 bit 的错误（毕竟误码率小于 50%）的情况下，仍然可以判断 c_W' 中有无水印（1 bit 信息），则 Eve 的攻击也将失败。而对于 Alice，为了阻止 Eve 的攻击，必须另寻隐蔽信道，比如可以采用 Cox 等提出的选择 DCT（discrete cosine transformation 离散余弦变换）域的中低频作为隐蔽信道，实现隐蔽通信。

基于以上分析，这里给出一些有益于数字水印嵌入、检测及攻击的指导性思想。我们认为：

（1）对于发送方 Alice，应该是努力扩大 H_{i_0} 的

范围，从而增加隐蔽信道的选择范围。

(2) 对于发送方 Alice，在保证嵌入水印后的图像 c_W 处于图 2 的 7 区的前提下，应该找到其中信道容量最大的嵌入方式，以最大化隐蔽信道的通信能力。

(3) 对于接收方 Bob，他既可以通过研究图像信道的性质以提高其检测性能，也可以针对特定的隐蔽水印信道的性质设计相应的检测器以阻止 Eve 的攻击。

(4) 对于攻击方的 Eve，如何正确识破 Alice 定义的水印不可感知性，从而找到隐蔽的水印信道是实现有效攻击的重要方法。

5　总　结

本文充分考虑了水印系统的不可感知性要求，得出了水印通信其实是一个在隐蔽信道上的一对一的通信过程的结论。分析了数字水印系统中发送者、攻击者及接收者之间的攻防关系，得出了一些有益于数字水印嵌入、检测及攻击的指导性思想。

参考文献：

[1] TIRKEL A Z, RANKIN G A, van SCHYNDEL R M, et al. Electronic watermark // Digital Image Computing, Technology and Applications (DICTA'93) [C]. Sidney: Macquarie University, 1993: 666－673.

[2] COX I, KILLIAN J. Secure spread spectrum Watermarking for Mulitmedia[J]. IEEE Trans on Image Proc, 1997, 6(12): 1673－1687.

[3] BARNI M, BARTOLINI F, CAPELLINI V, et al. ADCT-domain system for robust image watermarking[J]. Signal Process, 1998, 66: 357－372.

[4] PETITCOLAS F A P, ANDERSON R J, KUHN M G. Information hiding —— a survey[J]. Proc IEEE, 1999, 87(7): 1062－1078.

[5] COX I J, MILLER M L, McKELLIPS A L. Watermarking as communications with side information[J]. Proc IEEE, 1999, 87: 1127－1141.

[6] MOULIN P, O'SULLVAN J A. Information-theoretic analysis of watermarking // Proc Int Conf On Ac, Sp And Sig, Proc (ICASSP) [C]. Turkey: Istanbul, 2000.

Analysis of the Attack and Defend in Digital Watermarking

SUN Shui fa, QIU Pei liang, ZHANG Hua xiong

(1. Department of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China;
2. College of Information and Electronics, Zhejiang University of Science and Technology, Hangzhou 310023, China)

Abstract: The communication model of digital watermarking is investigated from the view of the imperceptivity of digital watermarking. The concepts of image space, perceptual space and watermark space is proposed and the image threshold, perceptual threshold and watermark threshold are defined on the three spaces. The relationship among the sender, the attacker and the receiver is analyzed. The results show that the key of digital watermarking is to find a covert channel, which is similar to the steganography. An information-theoretic model of digital watermarking is presented and one can use it to describe other information hiding technology, such as steganography. The model is analyzed and an example is given. Some constructive ideal for the sender, the attacker and the receiver is given.

Key words: digifal watermarhing; information aiding; coner chanels; informatim-theoretic model