Chapter 1.32
Data Hiding in Document Images

Minya Chen
Polytechnic University, USA

Nasir Memon
Polytechnic University, USA

Edward K. Wong
Polytechnic University, USA

ABSTRACT

With the proliferation of digital media such as images, audio, and video, robust digital watermarking and data hiding techniques are needed for copyright protection, copy control, annotation, and authentication of document images. While many techniques have been proposed for digital color and grayscale images, not all of them can be directly applied to binary images in general and document images in particular. The difficulty lies in the fact that changing pixel values in a binary image could introduce irregularities that are very visually noticeable. Over the last few years, we have seen a growing but limited number of papers proposing new techniques and ideas for binary image watermarking and data hiding. In this chapter we present an overview and summary of recent developments on this important topic, and discuss important issues such as robustness and data hiding capacity of the different techniques.

INTRODUCTION

Given the increasing availability of cheap yet high quality scanners, digital cameras, digital copiers, printers and mass storage media the use of document images in practical applications is becoming more widespread. However, the same technology that allows for creation, storage and processing of documents in digital form, also provides means for mass copying and tampering of documents. Given the fact that digital documents need to be exchanged in printed format for many practical applications, any security mechanism for protecting digital documents would have to be compatible with the paper-based infrastructure. Consider for example the problem of authentication. Clearly an authentication tag embedded in the document should survive the printing process. That means that the authentication tag should be embedded inside the document data rather than appended to the bitstream representing the document. The reason is that if the authentication tag is appended to the bitstream, a forger could easily scan the document, remove the tag, and make changes to the scanned copy and then print the modified document.

The process of embedding information into digital content without causing perceptual degradation is called *data hiding.* A special case of data hiding is *digital watermarking* where the embedded signal can depend on a secret key. One main difference between data hiding and watermarking is in whether an active adversary is present. In watermarking applications like copyright protection and authentication, there is an active adversary that would attempt to remove, invalidate or forge watermarks. In data hiding there is no such active adversary as there is no value associated with the act of removing the hidden information. Nevertheless, data hiding techniques need to be robust against accidental distortions.

A special case of data hiding is *steganography* (meaning *covered writing* in Greek), which is the science and art of secret communication. Although steganography has been studied as part of cryptography for many decades, the focus of steganography is secret communication. In fact, the modern formulation of the problem goes by the name of the *prisoner's problem.* Here Alice and Bob are trying to hatch an escape plan while in prison. The problem is that all communication between them is examined by a warden, Wendy, who will place both of them in solitary confinement at the first hint of any suspicious communication. Hence, Alice and Bob must trade seemingly inconspicuous messages that actually contain hidden messages involving the escape plan. There are two versions of the problem that are usually discussed—one where the warden is *passive,* and only observes messages, and the other where the warden is *active* and modifies messages in a limited manner to guard against hidden messages. The most important issue in steganography is that the very presence of a hidden message must be concealed. Such a requirement is not critical in general data hiding and watermarking problems.

Before we describe the different techniques that have been devised for data hiding, digital watermarking and steganography for document images, we briefly list different applications that would be enabled by such techniques.

1. **Ownership assertion:** To assert ownership of a document, Alice can generate a watermarking signal using a secret private key, and embed it into the original document. She can then make the watermarked document publicly available. Later, when Bob contends the ownership of a copy derived from Alice's original, Alice can produce the unmarked original and also demonstrate the presence of her watermark in Bob's copy. Since Alice's original is unavailable to Bob, he cannot do the same provided Alice has embedded her watermark in the proper manner (Holliman & Memon, 2000). For such a scheme to work, the watermark has to survive operations aimed at malicious removal. In addition, the watermark should be inserted in such a manner that it cannot be forged, as Alice would not want to be held accountable for a document that she does not own (Craver et al., 1998).
2. **Fingerprinting:** In applications where documents are to be electronically distributed over a network, the document owner would like to discourage unauthorized duplication and distribution by embedding a distinct watermark (or a fingerprint) in each copy of the data. If, at a later point in time, unauthorized copies of the document are