Recovering Invisible Image Watermarks from Images Distorted by the "StirMark" Algorithm

Gordon W. Braudaway
IBM Corporation, Thomas J. Watson Research Center
Yorktown Heights, NewYork 10598/USA
braud@watson.ibm.con

Abstract

"StirMark" is an image distorting algorithm developed at Cambridge University that is intended to attack and obliterate robust invisible image watermarks. The algorithm modifies a watermarked image by geometric distortion so subtle that the modification is essentially unnoticeable to a human observer. However, its effect on imbedded robust invisible watermarks can be devastating, successfully rendering them undetectable. In this paper, a countermeasure to the StirMark algorithm will be presented. It involves first detecting the presence of distortion in the distorted image, then measuring the magnitude and type of distortion, and finally removing the measured distortion thus restoring image geometry. Using a robust invisible watermarking method presented previously by the author, the results of removal of StirMark distortions will be demonstrated by showing successful extraction of an imbedded watermark from a realigned image.

Introduction

With the development of means of production and circulation of digital images, and the means of imbedding robust invisible watermarks into them ostensibly to convey image ownership, there is now financial incentive to attack an imbedded watermark and to attempt to render it undetectable. In nearly all invisible watermarking methods, pixel locations in a watermarked image can be presumed to correspond to those in an unmarked original image. A robust invisible watermarking method [1] previously presented by the author will be used as an example watermarking method. In this example method, the robust watermark is imbedded by altering only the values of the pixel components of the original image, not their geometric positions.

Very effective methods for attacking imbedded watermarks rely on constructing a new image, called a distorted image, that is derived from a watermarked image. One of the better known methods of doing this is called "StirMark." [2] Pixels in the distorted image are placed at subtly distorted positions relative to those in the watermarked image. All pixel values in the distorted image are determined by two-dimensional interpolation among enclosing pixel values in the watermarked image. Generally speaking, no constraints can be placed on the type of pixel position distortion that an attacker might choose to use, other than that the distorted image must not have obvious artifacts of the distortion process or appear to be a caricature of the watermarked image. This is a subjective measure, but it does require any linear or nonlinear distortion methods used to produce smooth and relatively small position distortions. The human visual system, as a qualitative measuring device, can be relied upon to readily detect excessive distortion. Hence, the only limit placed on any method of attack by the method of defense presented here is that the distortion of pixel positions the attacking method produces are generally small, so as to be essentially unobjectionable and casually unnoticeable to a human observer.

A Method of Defense

A method of defense against pixel-position distortion types of attacks, of which StirMark is an excellent example, is the subject of this paper. In the following discussion, an undistorted reference image is needed relative to which measurements of distortion can be made. For this purpose, either the original unmarked image or an invisibly watermarked copy of the original image can serve equally well as the reference image. The method requires first, a determination of the existence of pixel position distortion; second, measurement of the amount of pixel position distortion of three or more image features in the distorted image relative to corresponding features in the reference image; third, based on these measurements, calculation of coefficients of a pixel relocating equation that can specify an approximate position distortion for every pixel in the distorted image; and finally a pixel repositioning technique that can remove the measured distortion from the distorted image, thus approximately realigning it with its corresponding reference image. Once realignment of the distorted image with the reference image is achieved, ordinary watermark detection methods are used to detect the imbedded watermark. In fact, detection of an imbedded watermark from a realigned image will be the measure of the success of this defense method.

An Example Semiautomatic Embodiment

In an example semiautomatic embodiment, determination of the presence and measurement of the amount of pixel

position distortion can be accomplished as follows. First, if the presumed distorted image is not the same size as the reference image, it is made so by shrinking or enlarging the distorted image. Then the reference image and distorted image are each reduced to separate monochrome images, if they are not already monochrome images. A composite color image is constructed from the two monochrome images. The composite image has the monochrome reference image as its green color plane and the monochrome distorted image as its red color plane. If the composite image contains only shades of yellow (the sum of red and green) with no visible red and green fringes, the two images are correctly aligned and the presumption of distortion in the distorted image is not confirmed. Otherwise, if red and green fringes are evident, the two images are not aligned and restoration needs to be done.

Figure 1 is a monochrome copy of a composite color image composed as just described. The distorted image used was created by the StirMark algorithm and a monochrome copy of it is the red plane of the composite image. The green plane of the composite image is a monochrome copy of the reference image. (The reference image is the watermarked image used as input to the StirMark algorithm; it could also have been the unmarked original image instead.) Referring to Figure 1, it can be seen that light gray and dark gray fringes do exist. The light gray fringes correspond to green and the darker gray fringes correspond to red. The presence of fringes gives clear evidence that the distorted image is, in fact, distorted. By enlarging the composite image, using an image editor such as Adobe's Photoshop®, it is possible to record the horizontal and vertical pixel coordinates closest to corresponding features in the two images. A feature in this example might be the tip of a leaf on the tomato, the tip of a scratch on the cucumber, a blemish on the apple, a speck of dust on a napkin, or so on. The horizontal and vertical pixel coordinates closest to the n-th green (light gray) feature fringe will be referred to as x_n and y_n, respectively, and the coordinates of a pixel closest to a corresponding red (dark gray) feature fringe will be referred to as u_n and v_n.

Figure 1. A monochrome copy of the two-colored composite image. The light gray and dark gray fringes are not shadows, but are actually red and green fringes showing mis-registration of the distorted image relative to the reference image.

Due care must be exercised in selecting the red and green pixel positions of each unaligned feature. If the surrounding background is darker than the feature, the red and green colors fringes have a true representation. However, if the surrounding background is brighter than the feature, the colors fringes are reversed, red for green and green for red.

General equations can now be written that relate the coordinates of pixels in the reference image to the coordinates of corresponding pixels in the distorted image. The equations are:

$$x_n = \mathbf{a}\, u_n + \mathbf{b}\, v_n + \boldsymbol{c}$$

$$y_n = \mathbf{d}\, u_n + \mathbf{e}\, v_n + \boldsymbol{f}$$

Rewriting the equations in vector-matrix form gives:

$$\mathbf{X} = \mathbf{UA}$$

where:

$$\mathbf{X} = \begin{vmatrix} x_1 \\ y_1 \\ x_2 \\ y_2 \\ \cdots \\ x_n \\ y_n \end{vmatrix}, \mathbf{U} = \begin{vmatrix} u_1 & v_1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & u_1 & v_1 & 1 \\ u_2 & v_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & u_2 & v_2 & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ u_n & v_n & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & u_n & v_n & 1 \end{vmatrix}$$

$$\mathbf{A} = \begin{vmatrix} \mathbf{a} \\ \mathbf{b} \\ \mathbf{c} \\ \mathbf{d} \\ \mathbf{e} \\ \mathbf{f} \end{vmatrix}$$

The coefficients, **A**, can be evaluated, in a least-squares sense, for any value of *n* that is greater than or equal to three by using the following equation, providing the *n* pixels chosen do not lie in a straight line:

$$\mathbf{A} = [\mathbf{U}^{\mathrm{T}}\mathbf{U}]^{-1}\, \mathbf{U}^{\mathrm{T}}\mathbf{X}$$

where $\mathbf{U}^{\mathrm{T}}$ is the matrix transpose of **U**. Once the coefficients are solved for, the approximate undistorted pixel coordinates, *x* and *y*, for any values of the distorted pixel coordinates, *u* and v, are:

$$x = \mathbf{a}\, u + \mathbf{b}\, v + \mathbf{c}$$

$$y = \mathbf{d}\, u + \mathbf{e}\, v + \mathbf{f}$$

These equations, as a pair, are called the interpolation equations.

Realignment of the Distorted Image

Once the undistorted pixel coordinates of every pixel in the distorted image can be evaluated, a realigned image can be constructed. The undistorted pixel coordinates, in general, will not be distinct integers. One means of determining pixel values for the realigned image is to perform a two-dimensional interpolation using the pixel values from the distorted image that surround it. However, for the assumed small distortions, a simpler means can be used. The interpolation equations can be rewritten in vector-matrix form as:

$$\begin{vmatrix} x \\ y \end{vmatrix} = \begin{vmatrix} \mathbf{a} & \mathbf{b} \\ \mathbf{d} & \mathbf{e} \end{vmatrix} \begin{vmatrix} u \\ v \end{vmatrix} + \begin{vmatrix} \mathbf{c} \\ \mathbf{f} \end{vmatrix}$$

Ignoring terms of second order, the matrix of four coefficients can be factored into the product of four primitive matrices, as follows:

$$\begin{vmatrix} \mathbf{a} & \mathbf{b} \\ \mathbf{d} & \mathbf{e} + \mathrm{O}(2) \end{vmatrix} =$$

$$\begin{vmatrix} 1 & 0 \\ 0 & \mathbf{e} \end{vmatrix} \begin{vmatrix} \mathbf{a} & 0 \\ 0 & 1 \end{vmatrix} \begin{vmatrix} 1 & 0 \\ \mathbf{d/e} & 1 \end{vmatrix} \begin{vmatrix} 1 & \mathbf{b/a} \\ 0 & 1 \end{vmatrix}$$

The four primitive matrices, right to left, represent the primitive image manipulating operations of horizontal shearing, vertical shearing, horizontal scaling, and vertical scaling. Those four primitive operations, applied to the distorted image in the sequence specified, will restore the distorted image to the a close approximation of the reference image. The restoration will be called the realigned image. It is easily verifiable that the four primitive matrices can be rearranged into several different sequences that will produce essentially the same result.

One operation remains in restoring the distorted image to a very close approximation of the reference image. In the process of restoring the distorted image, it is possible that pixels at the edges of the realigned image will have been lost, and the center of the realigned image may be horizontally and vertically offset from that of the reference image. This is predicted by non zero values of coefficients **c** and **f** in the interpolation equations. Rather than use those coefficients as the offset of the realigned image relative to the reference image, a two-dimensional discrete cross-correlation surface is computed employing two-dimensional discrete Fast Fourier Transforms (FFT's), and the interpolated horizontal and vertical offsets of the peak value of the cross-correlation surface relative to its origin are used instead. This is a somewhat more reliable means of finding the best offset values if the original distortions were nonlinear, since the method described above presupposes them to be linear. For images distorted by StirMark algorithm, it makes no practical difference.

Detection of the Invisible Robust Watermark after Image Restoration

The method described above has been tested by application to a StirMark distorted image. The image used was the same that used in Figure 1. First the image was watermarked and a baseline watermark extraction was performed. The baseline watermark extracted is shown in Figure 2. The watermarked image was then distorted by the StirMark algorithm, com-

pressed 15:1 using JPEG compression, and a watermark extraction was attempted from the distorted image. The result is shown in Figure 3. The watermark has been effectively obliterated by Stirmark algorithm. Finally, the distorted image was realigned by the method described above, and the watermark was extracted from the realigned image. The result, shown in Figure 4, indicates that the watermark was successfully extracted from the realigned image with little deterioration caused by the extensive processing of StirMark distortion, JPEG compression, and subsequent restoration.

*Figure 2. A "baseline" robust invisible watermark extracted from the **watermarkedimage** prior to its being processed by the "Stirmark" algorithm.*

*Figure 3. A robust invisible watermark extracted from the **watermarked image** after it was processed by the "Stirmark" algorithm and JPEG compressed 15:1, showing complete obliteration of the watermark.*

*Figure 4. A robust invisible watermark extracted from a **realigned image**, after restoration by the process described, showing that relatively little damage was done by the combined "Stirmark" distortion, JPEG compression, and restoration processes.*

Concluding Remarks

The image restoration process detailed here has shown itself to be remarkably effective at restoring images distorted by the StirMark algorithm. The restoration process is quite general, in that it does not depend on any inside information concerning the types of distortion used by StirMark. In particular, using arguments from basic calculus, it is irrelevant whether linear or nonlinear methods are used for the distortion, as long as any nonlinear distortions are relatively small. The restoration method should, therefore, be applicable to watermarking attacking methods other than StirMark that rely on subtle distortion of image pixel locations.

The entire restoration process can be repeated iteratively, if necessary, by substituting the previously realigned image for the distorted image before each subsequent iteration. For StirMark, no iteration was required, but for other attacks with more nonlinear distortion, iteration might be needed.

The invisible watermarking method referenced[1] has been demonstrated to imbed watermarks that survive and remain extractable after a watermarked image is printed and rescanned. It is not hard to find subtle, unintended distortions in the printing and rescanning process that might damage an imbedded watermark, although this has not been the case so far. The restoration method described here could also be applied to printed and rescanned images in the same manner it has been applied to images distorted by the StirMark algorithm, and thus it remains a powerful technique in the watermark extraction arsenal.

An automatic restoration method based on the principles described above, but not requiring human assistance with an image editor, is the subject of continuing research at the IBM T.J. Watson Research Center.

References

1. Gordon W. Braudaway, Protecting Publicly-Available Images with an Invisible Image Watermark, *Proc. of the IEEE ICIP'97, Vol. I*, Santa Barbara, CA, October 26-29, 1997, pp. 524-527.
2. Fabien A.P. Petitcolas, Ross J. Anderson, Markus G. Kuhn: Attacks on Copyright Marking Systems, in David Aucsmith (Ed.): *Information Hiding, Second International Workshop, IH'98*, Portland, Oregon, April 15-17, 1998, Proceedings, LNCS 1525, Springer-Verlag, ISBN 3-540-65386-4, pp. 219-239.