

Improved Key Management for Digital Watermark Monitoring

Volker Roth and Michael Arnold

Fraunhofer Institute for Computer Graphics
Dept. Security Technology
Rundeturmstrasse 6, D-64283 Darmstadt, Germany

ABSTRACT

In this article we propose content based retrieval techniques as means of improving the key management used for digital watermark monitoring. In particular, we show how keys for watermark monitoring can be used on a per media item basis (rather than one secret key for all works copyrighted by a single owner), while retaining a high probability of successful spotting.

Keywords: Digital watermarking, content based retrieval, monitoring, intellectual property protection, cryptography

1. INTRODUCTION

Digital Watermarking is a means for spotting illicit use of copyrighted material in mass media such as television, radio, and the Internet. The general principle is as follows: the copyright owner embeds secret copyright labels into the media by means of digital watermarking. Either the copyright owner himself or his contractor monitors custom distribution channels for copyrighted material that is not licensed for the detected use. Ideally, the copyrighted material is marked with a general copyright label as well as a label that identifies the customer who licensed the material in question. In a general setting, copyright owners license multiple different works. This results in one secret key being used to embed the copyright labels in all copyrighted works of that owner, such that the monitoring process can be applied uniformly to all copyrighted works.

However, this approach has a serious drawback – once the secret key is known, the copyright labels in all the works copyrighted by that owner can be removed easily. Unfortunately, there may be situations, where the owner of a copyright is forced to disclose his key to a third party, for instance in the course of an ongoing dispute over an alleged license violation.

One solution to this problem is to use unique keys for each copyrighted work. Consequently, the monitoring process must have efficient means of mapping a monitored work to the key that was used for watermark embedding (we refer to this as the *identification problem*). The mapping must be robust with respect to the same criteria that apply to the robustness of the digital watermark itself.

One technology that can be brought to bear on the identification problem is *content-based retrieval*.¹ In the monitoring process, the content-based feature extraction and comparison algorithms must discriminate between the original and a possibly manipulated watermarked copy, and two media items that are unrelated to each other. Since media ultimately serve humans, we conjecture that any manipulation of a media item that breaks the human perception of similarity renders the media item useless in its original sense, or is a copyrightable work in itself.

In our paper, we discuss the application of content-based retrieval technology to the identification problem in watermark monitoring, and propose a key schedule for watermark embedding where copyright ownership of a work can be demonstrated without putting the copyright labels of other works at risk. The usage of content-based retrieval methods to

Copyright 2002 Society of Photo-Optical Instrumentation Engineers. This paper was published in Proc. SPIE 2002, San Jose, CA, USA, and is made available as an electronic reprint with permission of SPIE. One print or electronic copy may be made for personal use only. Systematic or multiple reproduction, distribution to multiple locations via electronic or other means, duplication of any material in this paper for a fee or for commercial purposes, or modification of the content of the paper are prohibited.

solve the identification problem can be combined with *nonblind watermarking*² in order to increase the reliability of the watermark recovery process.

In Sect. 2 we describe the media preparation and initialization, as well as necessary assumptions and requirements. The actual watermark monitoring process is the subject of Sect. 3. We also discuss the anticipated advantages of combining watermark monitoring with content-based retrieval. Some brief conclusions are given in Sect. 5.

2. MEDIA PREPARATION

For illustration we use digital images as an example media type, although the general scheme can be applied to any media type for which suitable watermarking and content-based retrieval algorithms are known. Our description is based on the definitions given below.

Let A be a media agency. Whenever a new image I is added to the stock of the agency, a secret key k_I is randomly generated for that image. The agency keeps a database that stores annotations of each image. For a known image I , that is in stock, its corresponding record can be retrieved efficiently. Particularly k_I is stored in that record.

Let \mathcal{M} be a message space, $\mathcal{K} \subseteq \mathcal{M}$ be the key space, \mathcal{I} be the set of images, and let $\mathcal{F}_1 \subset \{f \mid f : \mathcal{I} \rightarrow \mathcal{I}\}$ be a family of transformations that includes the identity transformation. We define a *nonblind* watermark embedding function e and watermark detection function d as follows:

$$\begin{aligned} e & : \mathcal{I} \times \mathcal{K} \times \mathcal{M} \rightarrow \mathcal{I} \\ d & : \mathcal{I} \times \mathcal{I} \times \mathcal{K} \times \mathcal{M} \rightarrow \{true, false\} \end{aligned}$$

such that

$$\forall (k \in \mathcal{K}, m \in \mathcal{M}, I \in \mathcal{I}, f \in \mathcal{F}_1) : d(f(e(I, k, m)), I, k, m) = true$$

In other words, e and d are 1-bit nonblind watermark embedding and detection algorithms, and the watermarking scheme is robust against the transformations in \mathcal{F}_1 . Furthermore, e shall fulfill the general requirements for watermarking functions as proposed e.g., in Ref. 3. Whenever a customer B wishes to license an image I , the agency computes $I' = e(I, k_I, m)$ and hands out I' to the customer. The watermark m carries the unique *copyright marker* of the agency, which should be registered officially (similar to a trademark).

3. MONITORING

We assume that the watermark monitoring process comes across a watermarked image $I'' = f(I')$, $f \in \mathcal{F}_1$. Obviously, the monitoring process can detect the copyright marker only if the original image I and the secret key k_I , that was used to embed the marker, are known. This requires that I'' is mapped onto the original image I from which it was derived. The mapping function is implemented by means of *content-based image retrieval* (CBIR), a term whose earliest use in literature, according to Eakins,¹ seems to have been by Kato.⁴

CBIR systems are generally based on a *feature extraction* mechanism and a *distance metric* that is defined on the feature domain. The feature extraction mechanism takes an image as its input and outputs a compact representation of the salient visual features of that image, also called a *feature vector*. The distance metric gives a measure of similarity of two feature vectors. More formally, let Ω be the feature domain, $\xi : \mathcal{I} \rightarrow \Omega$ the feature extraction, $\mathcal{S} = \{\omega_1, \dots, \omega_N\}$ a *feature dataset* whose elements are the feature vectors of the images in stock, and $\delta : \Omega \times \Omega \rightarrow \mathbb{R}^+ \cup \{0\}$ the metric function.

A search for the original image that corresponds to a given query image I'' with feature vector $\omega = \xi(I'')$ can then be formulated as a *K nearest neighbors query* on the feature dataset \mathcal{S} , denoted $knn_{\mathcal{S}}(\omega, K)$, where

$$knn_{\mathcal{S}}(\omega, K) \subseteq \mathcal{S} \tag{1}$$

$$|knn_{\mathcal{S}}(\omega, K)| = \min(K, N) \tag{2}$$

$$\forall \omega_i \in knn_{\mathcal{S}}(\omega, K) \neg \exists \omega_j \in \mathcal{S} \setminus knn_{\mathcal{S}}(\omega, K) : \delta(\omega, \omega_j) < \delta(\omega, \omega_i) \tag{3}$$

Efficient algorithms for a K nearest neighbors search in vector spaces of a high dimension is still a subject of ongoing research, and depend heavily on the organization of the index.⁵ A trivial implementation compares the query feature

vector with all elements of the feature dataset. The feature vectors of the images in the stock are usually computed a priori. Ideally, there is only one image in the stock whose feature vector has the minimum distance to the feature vector of the query image, hence

$$\xi(I) \in \text{knn}_{\mathcal{S}}(\xi(f(e(I, k, m))), K) = \text{knn}_{\mathcal{S}}(\xi(I''), K), \quad K \text{ is "small"}$$

for all $f \in \mathcal{F}_2$ and $\mathcal{F}_2 \subset \{f \mid f : \mathcal{I} \rightarrow \mathcal{I}\}$. In other words, \mathcal{F}_2 is the family of image transformations against which the CBIR algorithm is robust. We expect that for well-chosen CBIR algorithms $|\mathcal{F}_2| > |\mathcal{F}_1|$ holds, and that $|\mathcal{F}_2 \setminus \mathcal{F}_1|$ is “big.”

The reason why we believe this is the following: watermarking algorithms are designed to be visually imperceptible. Therefore, only details of an image are modified by the watermark embedding process. On the other hand, CBIR algorithms are designed to extract only the most salient visual features of an image, omitting details in the process. An image transformation that “breaks” the feature extraction and distance metric likely breaks the human perception of visual similarity as well.

For instance, CBIR algorithms based on color histograms, such as the *Color Coherence Vector* algorithm⁶ (CCV), are robust against transformations such as shearing, scaling, and rotations by small angles. Even more important, we can expect that the CCV algorithm is fairly robust against general mesh warping (which is used e.g., by the StirMark⁷ attack), as long as the geometric distortions are not visually perceptible.

The characteristics of the scheme that we propose can be exploited in order to improve watermark monitoring at least fourfold:

1. The monitoring process can retrieve I_j and k_{I_j} for each $\omega_j \in \text{knn}_{\mathcal{S}}(\xi(I''), K)$, and use this information for watermark detection. This effectively turns a *nonblind* detection algorithm into a *blind* detection algorithm.
2. The images I_j for $\omega_j \in \text{knn}_{\mathcal{S}}(\xi(I''), K)$ can be used to invert geometric distortions in I'' , thus increasing the probability of finding a watermark.
3. A unique secret key can be used for each image. Therefore, a single leaked key does not compromise the security of the copyright markers that are embedded into other images of the same agency.
4. The watermark embedding and detection algorithms can be tailored to images of a particular class, thus becoming more robust. Each image in the stock can be annotated with the identifier of the detection algorithm that must be used for it.

Of course, these advantages are paid for with the added overhead of the monitoring process, that incurs due to the K nearest neighbors search. It remains to be investigated in which cases the proposed scheme outperforms the trival approach (e.g., testing for a watermark by trying all N keys of the images in the stock). The advantage to be able to tailor watermark embedding and detection to a particular image class is unaffected by these concerns. The general watermark monitoring process for a given candidate image is illustrated by Alg. 1.

4. EXPERIMENTS

We were interested to see how potential malicious image manipulations affected the query results of our example CBIR algorithm (the CCV⁶ algorithm). As a simple experiment, we generated about 50 distorted query images from a set of base images. All base images showed football scenes, and many images were rather similar to each other. The distortion that we applied, and to which we refer subsequently as β , consisted of the sequence of transformations given below:

1. horizontal flip;
2. cyclic horizontal roll by 1 pixel;
3. clockwise rotation by one degree;
4. resampling to 90% of the width and height of the original image.

Algorithm 1 The abstract monitoring algorithm.

```
1: {Input is the candidate image  $I''$ . The output is true if the copyright mark is embedded in  $I''$ .}
2:  $\omega = \xi(I'')$ 
3: {First, we attempt to detect the copyright marker using the keys of all matches in the  $K$  nearest
   neighbors search.}
4: for all  $j$  with  $\omega_j \in knn_{\mathcal{S}}(\omega, K)$  do
5:   if  $d(I'', I_j, k_{I_j}, m) = true$  then
6:     return true
7:   end if
8: end for
9: {If the copyright marker was not found then we try to invert any geometric distortions of the
   candidate image with respect to its  $K$  nearest neighbors and run the detection again.}
10: for all  $j$  with  $\omega_j \in knn_{\mathcal{S}}(\omega, K)$  do
11:   Determine inverse transformation  $f^{-1}$  based on  $I''$  and  $I_j$ .
12:   if  $d(f^{-1}(I''), I_j, k_{I_j}, m) = true$  then
13:     return true
14:   end if
15: end for
16: {If still no copyright marker was found but one of the  $K$  nearest neighbors of the candidate
   image is “very close” to the candidate image (e.g., has a distance smaller than a given  $\epsilon$ ) then
   a human should probably compare these images. In order to make denial of service (DoS)
   attacks harder, we introduce a probability for inspection that can be adapted to the load of the
   process.}
17: for all  $j$  with  $\omega_j \in knn_{\mathcal{S}}(\omega, K)$  do
18:   if  $\delta(\omega, \omega_j) < \epsilon$  then
19:     Schedule  $(I'', j)$  with probability  $p$  for concurrent human inspection.
20:   end if
21: end for
22: return false
```

Then we queried the original image database with each one of the distorted images, using the CCV algorithm, and measured the normalized distances of the best and second best matches. In all cases, the correct original image was identified as the best match. Figure 1 shows a plot of the distances of the best matches (denoted by circles) and the second best matches (denoted by cross hairs).

The plot clearly shows the correlation between the distorted query images and the originals from which they were derived. The average distance is scattered closely around 0.05. Two extreme cases can be identified. At query image 40, the best match (original image number 40) and second best match (original image number 39) are both very similar to the query image. The reason is that both originals have a high degree of similarity to each other in the first place.

The second interesting case is that of a query where the best match has a comparably high distance to the original image from which it was derived. This means that the distortion had a greater impact on the robustness of the CCV algorithm due to intrinsic features of the original image. One such example is found at image number 37. However, the distance to the second best match is even greater.

Although the results are encouraging, care must be taken when considering CBIR algorithms. For instance, the CCV algorithm is not robust against changes in the brightness of images when it is computed on the RGB color model. Figure 2 illustrates the effects. For this test, we computed the query images simply by applying a gamma correction of 0.8 to the original images. Then we queried the original image database with each of the gamma corrected query images, and assigned the query image a rank that is equal to the position of its corresponding original in the sorted result set. In other words, a query image was assigned rank n if its corresponding original image was the n 'th best match in the result set. A

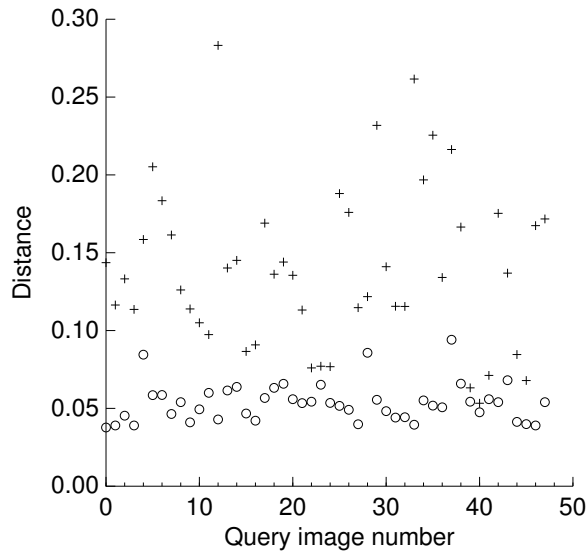


Figure 1. This plot shows the distances of the best matches (denoted by circles) and second best matches (denoted by cross hairs) to our distorted query images.

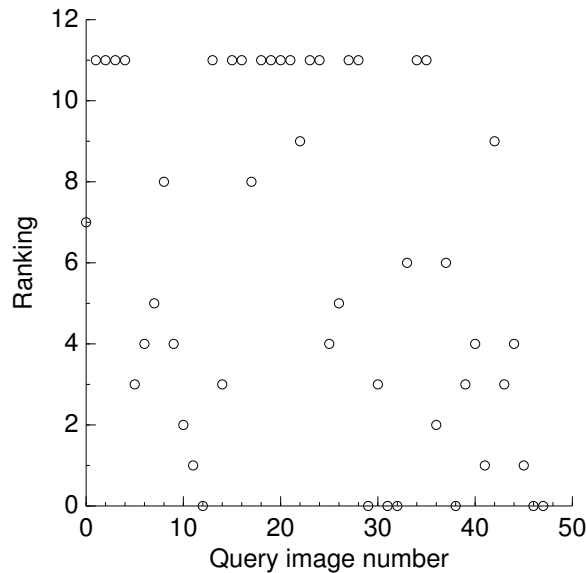


Figure 2. This plot shows the rankings of the query images, where the distortion was a 0.8 gamma correction. A ranking of n means that the original image of a given query image appeared at position n in the sorted set of results. A rank of 11 means that the original of a query image was not among the 10 best matches for that query image.

ranking of 11 denotes the “point at infinity.” This means that the original image for a given query image was not among the 10 best matches in the result set.

Hence, we have a case where some watermarking algorithms⁸ are fairly robust against gamma correction (denoted γ), while the chosen CCV based CBIR algorithm is not. On the other hand, the CCV algorithm, when computed on the RGB color model, is robust against the distortion β where most watermarking algorithms are not. Sticking to our notation, this means that for our given combination of CBIR and watermarking algorithms probably the following properties hold:



Distorted query images.



Best matches (original images).



Second best matches.

Figure 3. This figure shows the query image, original, and second best match for two extreme cases. The left column shows a case where the best match and second best match are extremely close in terms of CCV distance. The right column shows a case where the best match has a comparably huge distance to the distorted query image.

$\beta \in \mathcal{F}_2, \beta \notin \mathcal{F}_1, \gamma \in \mathcal{F}_1, \gamma \notin \mathcal{F}_2$. Furthermore, we may safely assume that $\beta \circ \gamma \notin \mathcal{F}_1 \cup \mathcal{F}_2$. One conclusion is that $\beta \circ \gamma$ poses a threat to the instance of the watermark monitoring process that we discussed in our article. The general question that might be asked is which CBIR algorithms can be applied best in conjunction with a given watermarking algorithms such that the monitoring process is “optimal.”

5. CONCLUSIONS

In this article, we proposed to use *content based retrieval* as a means to improve the monitoring of digital media for watermarks. Content based retrieval maps a candidate media item to likely originals based on salient features of the candidate item. Whereas digital watermarking aims to be imperceptible, hence manipulates only details, content based retrieval is designed to be robust against imperceptible distortions. Both techniques complement each other nicely.

The principal benefits that can be expected from that approach are the following: a) *nonblind* detection algorithms can be turned into *blind* ones; b) likely originals of a candidate image can give hints to be used for inverting transformations that obscure watermarks embedded in the candidate item; c) a unique key can be used for watermark embedding on a per media basis, hence a single leaked key does not compromise other media items; d) the watermark embedding and monitoring algorithm can be tailored to a specific domain of media items, which can improve the robustness and imperceptibility of the watermarking. An abstract monitoring algorithm is given that illustrates these benefits.

6. ACKNOWLEDGMENTS

We would like to thank Stephan Volmer for his helpful review of the initial draft version of this paper, and his suggestions for improvement.

REFERENCES

1. J. Eakins and M. Graham, "Content-based image retrieval." Report to JISC Technology Applications Programme, January 1999. Available at URL <http://www.unn.ac.uk/iidr/report.html>.
2. S. Katzenbeisser and F. A. P. Petitcolas, *Information hiding techniques for steganography and digital watermarking*, Artech House, Inc., 685 Canton Street, Norwood MA 02062, 2000.
3. I. D. Bramhill and M. R. C. Sims, "Challenges for copyright in a digital age," *BT Technol J* **15**, pp. 63–73, April 1997.
4. T. Kato, "Database architecture for content-based image retrieval," in *Image Storage and Retrieval Systems*, A. A. Jambardino and W. R. Niblack, eds., *Proc. SPIE* **1662**, pp. 112–123, 1992.
5. S. Volmer, "Buoy Indexing of Metric Feature Spaces for Fast Approximate Image Queries," in *Proc. Eurographics 2001 Workshop on Multimedia*, pp. 121–130, (Manchester, UK), September 2001.
6. G. Pass, R. Zabih, and J. Miller, "Comparing images using color coherence vectors," in *Proc. ACM Conference on Multimedia*, (Boston, Massachusetts, U. S. A.), November 1996.
7. F. A. P. Petitcolas, R. J. Anderson, and M. G. Kuhn, "Attacks on copyright marking systems," in *Second International Workshop on Information Hiding, 14–17 April, 1998, Portland, Oregon, USA*, D. Aucsmith, ed., *Lecture Notes in Computer Science* **1525**, pp. 219–239, Springer-Verlag, (Berlin, Germany / Heidelberg, Germany / London, UK / etc.), 1998.
8. L. Piron, M. Arnold, M. Kutter, W. Funk, M. Boucqueau, and F. Craven, "OCTALIS benchmarking: Comparison of four watermarking techniques," in *Security and Watermarking of Multimedia Contents*, P. W. Wong and E. J. Delp, eds., *Proc. SPIE* **3657**, pp. 240–250, SPIE – The International Society for Optical Engineering, SPIE, (San Jose, California), January 1999. ISBN 0-8194-3128-1.