Digital Watermarking Techniques and Security Issues

Jordi Nin and Sergio Ricciardi
Department of Computer Architecture, Technical University of Catalonia - BarcelonaTECH (UPC)
nin@ac.upc.edu, sergio.ricciardi@ac.upc.edu

***Abstract*—Digital watermarking is the process of embedding information into a noise-tolerant digital signal such as image or audio data to easily identify the copyright ownership of the media. Such information is embedded for many different purposes, such as copyright protection, source tracking, piracy deterrence, etc., and therefore it shall be embedded in a way that makes it difficult to be removed. There are an extensive literature about watermarking algorithms and methods as well as possible attack techniques. In this work we collect a part of this vast literature in order to make easy for a non-expert reader about watermarking to have a high-level overview on new trends a technologies related to image and multimedia watermark algorithms and attacks.**

***Keywords*-Watermarking; fingerprinting; digital right management systems; image security; watermark attack techniques**

I. Introduction

Digital media has many advantages over analog media [1], however the possibility of unlicensed duplication and dissemination of copyrighted material poses a hazard to traditional business models [2], [3]. Two complementary techniques have been applied to address this problem: encryption and watermarking. Encryption techniques can be used to protect the data when it is being delivered from the sender to the receiver. The receiver decrypts the data and obtain the original copy. Complementary to encryption, invisible watermarking [4] can embed a secret (possible imperceptible) signal, called *watermark*, such that it can not be (easily) extracted but that can be easily read and employed in many different applications [5].

Each watermark signal is application-specific; nevertheless, general requirements for a watermark can be specified as it is described in [6]. A key feature of watermarking is the *perceptual transparency*. It refers to the fact that the embedding of a signal should not be perceptible to humans and not affect the quality of the underlying data [7].

Watermarks can be robust or fragile: a digital watermark is fragile if it fails to be detectable after the slightest modification, and is robust if it resists a designated class of transformation. Depending on the application requirements, a fragile or robust watermarking technique can be appropriate. If the desired behavior is the integrity proof (tamper detection) then a fragile watermark is enough; whereas, if a watermark is used to carry copyright notices and prevent unauthorized copies, it is important that it is robust, and can survive the many attacks that may be thrown at it to eschew theft detection and prosecution [3].

In this work, we describe most common watermarking application scenarios to illustrate that watermark algorithms are in the eye of the storm of most of the Internet security and Copyright problems. Then, we summarize the most common and well-know watermark methods giving a readable description and explaining their main advantages and drawbacks. Additionally, we provide a description of many possible attacks against watermarks.

II. Applications

Digital Watermarking describes methods and technologies that embed hidden information, for example a number or a string, in digital media, such as images, video, audio or any other kind of noise-tolerant digital signal such as multimedia data. Special softwares are available for embedding imperceptible information via subtle changes to the data of the original digital content [8]. Digital watermarks can be easily detected and read by computers, networks and a variety of digital devices, validating the original content and/or initiating or preventing actions.

Figure 1 illustrates the workflow of any watermark. Initially, the watermark is embedded inside the original image. Then, the watermarked imaged is copied and (or) distributed. After that, the image is often cropped, resized, compressed, etc. However the metadata within the watermark has to remain unchanged to allow *tracebility*.

In the following, some existing application areas are described together with the reference technologies, and case studies are presented, highlighting some of the most common real world scenarios. Most of the examples shown refers to the watermarking of digital images, but they are in general applicable to other media, such as audio or video streams.

A. Copyright protection

The first application area to which watermarking was employed is the copyright protection of digital media. In the digital world it is possible for almost anyone to duplicate or manipulate digital data without loosing quality. This has allowed previously unseen copyright infringement issues. Digital watermarking provides an added layer of security to the content protection chain to deter unauthorized use/duplication of content by embedding watermarks that

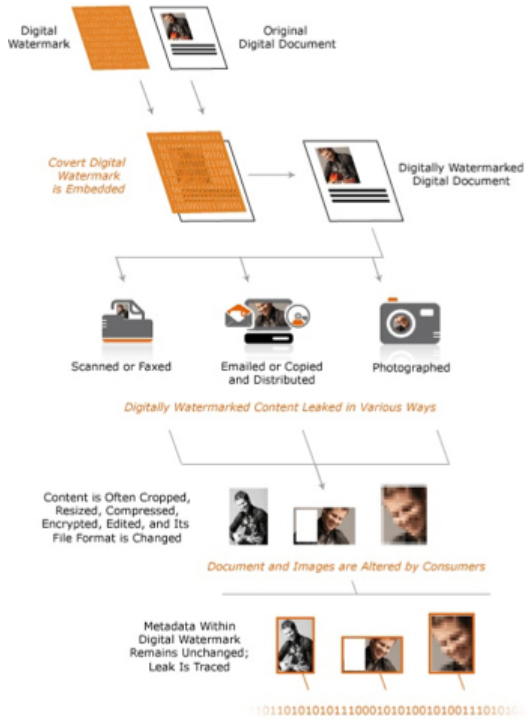

Figure 1. Workflow of Digital Watermarking.

identify the original media and the permitted uses of the content.

In such a scenario, devices read the watermark during playback or copying of the content. If the watermark indicates that the use is unauthorized, the playback or copying is prevented (other actions are also possible, e.g. the audio is muted), and an explanatory message may be displayed. Effective content protection helps content owners to protect audio, film and video entertainment content, communicate copyright ownership and usage rights of their content, protect it against common threats of piracy including camcorder recording, peer-to-peer file sharing, copying, format conversion, encoding and other forms of re-processing.

B. Content identification and management

Digital watermarking enables effective content identification by providing a unique digital identifier to all forms of media content in a way that persists with the content wherever it may travel. Digital watermarks are easily embedded into content without interfering with the consumer's enjoyment of it. It is imperceptible to humans, but easily detected and understood by computers, networks and a wide range of common digital devices. The watermark can carry such information, such as the owner identity, how it may be used or anything else the owner wants to convey. It can also trigger predefined actions, including linking to websites or other consumer experiences. Content identification helps:

- Consumers to find the content they are looking for, learn more about it, try it out, and locate where and how to purchase it;
- Copyright owners, brands and distributors to locate and learn about how, when and where content is being consumed and identify the source of leaks when confidential content inadvertently or intentionally makes its way onto the Internet.

C. Content filtering: triggering of actions and blocking

The data carried in the digital watermark can be rapidly cross correlated with other content or actions. On the one hand, a specific action or even piece of content can be triggered upon identification of the watermark, allowing customer interactivity. For instance, while watching a scene in a movie, a specific call to action (e.g., "Press the red button on your remote to find out more") could be triggered. Similarly, a specific and targeted advertisement could be triggered. Instead of a commercial appearing at regular times, the commercial could be triggered according to what content is being watched and at specific times within the content.

On the other hand, digital watermarks could be used for blocking specific contents. Upon recognition and identification of a particular situation, the content could be blocked. Such applications can prove extremely useful for the Internet, such as blocking a copyrighted piece of audio or video from being uploaded to a website. Additionally, to ensure child safety and prevent children from being exposed to adult content, specific rules could be set up by the parent to warn, restrict or completely block the viewing of such a content.

D. Online contents

In the corporate world, images, documents and video quickly spread through emails and across the world wide web. In the case of major brands, for instance, marketing departments must carefully manage the release of product launch materials and ensure that their sales channels are correctly using the right images at the right time. Internet search services are available that constantly crawl the web looking for uniquely watermarked content. Reports are then generated notifying the owner of where their content was found, allowing them to take actions deemed necessary. Once content is found, a wide range of automated actions or messages are available, from the classic "This content is available for licensing." to the more intimidating "This content is copyrighted; please remove it immediately".

E. Document and image security

A unique digital watermark can be easily embedded into each copy of a confidential document as they are being created and distributed. The data contained in the watermark can include who are the recipients of each copy so that any information that is inadvertently or intentionally leaked out

is easily traced back to the source. Additionally, companies can use network detectors and email filters to check for digital watermarks within documents and images, providing notification if an attempt is made at uploading to the web or forwarding in email outside the company. Similarly, watermark detectors can be included in various printers, scanners and other devices to check for watermarks in confidential documents that someone is attempting to copy. In this case the watermark can trigger an action, such as a *do not copy or scan*.

Therefore, document and image security helps to:

- Identify each copy of a confidential document and/or image with a unique digital identity;
- Trace back to the source of leaks if sensitive materials are distributed intentionally or inadvertently;
- Filter documents being uploaded to the web or forwarded in email to quickly identify confidential materials and stop distribution;
- Prevent the copying of confidential documents on copiers and/or scanners;

F. Forensics and piracy deterrence

Forensic watermark applications enhance a content owner's ability to detect and respond to misuse of its assets. Forensic watermarking is used not only to gather evidence for criminal proceedings, but also to enforce contractual usage agreements between a content owner and the people or companies with which it shares its content. It provides positive, irrefutable evidence of misuse for leaked content assets. Forensics and piracy deterrence helps:

- Create a powerful deterrence from leaking controlled content either maliciously or unintentionally;
- Gain visibility over where and how their content is being accessed without the need of a complete DRM system [10] to restrict access.

G. Communication of ownership and copyrights

Watermarks resides embedded in the content as it is forwarded and travels across the Internet, and can be detected at any point to determine the content's unique identity. Watermarks also survive many different file manipulations and transformations, unlike standard metadata that is often lost leaving the content "orphaned". Copyright communication can be used to:

- to ensure content owners that their ownership and contact information stays permanently attached to their content wherever it may travel and be accessed on the web or packaged media;
- better manage content through a range of automated remedies when unauthorized use is discovered, including device enforcement messages of copyright policies, take down notices or providing permission with proper attribution.

H. Mobile Experiences and Watermarking

The watermarks can be easily embedded into all forms of media content, including magazines, newspapers, packaging, posters, brochures and more. And, unlike 2D barcodes or QR codes that are being used in some mobile campaigns, digital watermarks are imperceptible to humans and do not take up precious space on printed materials, making the technology much more "brand friendly". The digital ID in the watermark can be matched to a URL in a backend database that is then returned to the consumer' phone. The opportunities and experiences enabled by the technology include proprietary content for paid subscribers, contests and promotions, video contents, games, discounts, etc.

I. Audience measurement

Digital watermarking embeds a unique identifier into media content not only prior to distribution, but also while being distributed, making content and corresponding broadcasters instantly identifiable. Using specialized software able to retrieve, analyze and report the data, digital watermarking allows the precise identification of content and broadcasters.

In an audience measurement application, the technology works by inserting digital data, imperceptible to the human ear, into each program's audio track. The digital ID contains information about the channel that broadcast the program, the airing time and, if relevant, a content identifier. Audiometers, installed in panelists' houses, read the data, collect the information and periodically send them to a central database for processing and accurate reporting. Audience measurement includes:

- Accuracy and detailed detection logs allows the reporting of the content being watch, channel, airing time and distribution network;
- broadcasting media such as radio, television, Internet video and podcast audience measurement applications;
- integration into legacy audiometer to minimize change for panelists and audience operators.

III. Algorithms

In general, a watermark can be embedded in spatial, transform, frequency or compressed domain of a media (e.g., an image).

Spatial domain techniques directly modulate the pixels of the image. In this approach [11], [12], [4], the pixel value of an image is modified to embed watermark information. A very simple approach refers to the bit-plane manipulation of the least significant bit (LSB). This technique offers easy and rapid decoding and consists of the embedding of the m-sequence on the LSB of the image data. This method is also used in steganography. Another method is the linear addition of the watermark to the image data. This method is more difficult to decode and offers inherent security. The decoding process requires the examination of the complete bit pattern and its current implementation must

therefore be performed offline (i.e., once received the entire image), which represents its principal drawback. The main problem found with adding the watermark is in retaining the dynamic range of the original image and the auto-correlation output. The watermark is robust to averaging, and potentially compatible with JPEG compression.

Transform domain techniques modify the discrete cosine transform (DCT), discrete wavelet transform (DWT), discrete fourier transform (DFT) or any other transformed coefficients. Transform domain techniques [1], [5] usually achieve better performance since the perceptual characteristics of images can be better utilized and the spread spectrum principles used in secure communications can be easily incorporated. Typically, transform domain systems perform the watermarking process independent of compression, although these two processes share some common features. Transform domain techniques embed watermarks with visually recognizable patterns in the images as a set of independent and identical distributed sequences drawn from a Gaussian distribution into the perceptually most significant frequency components of an image [13]. The embedding positions selectively modify the middle frequency of DCT of the images. The embedding and extracting methods of the DCT-based approach have been described [13]. On the other hand, several methods use the discrete wavelet transform (DWT) to hide data to the frequency domain to provide extra robustness against attacks.

Lossy compression is an operation that usually eliminates perceptually non-salient components of an image. If one wishes to preserve watermark in such an operation, the watermark must be placed in the perceptually significant region of the data. Most processing of this sort takes place in the human perceivable frequency domain. In fact, data loss usually occurs among the very low/high frequency components. Hence the watermark must be placed in the significant frequency components of the spectrum. In contrast to the spatial-domain-based watermarking, frequency domain-based techniques can embed more bits of watermark and are more robust to attack; thus, they are more attractive than the spatial domain-based methods [13] for multimedia watermarking.

Compressed domain techniques integrate the compression framework with watermarking by directly labeling the compressed (quantized) symbol streams. There is little loss in generality by assuming a compresses domain framework since compression is nearly ubiquitous for multimedia. This finds application for embedding watermark in video data. A real-time watermarking algorithm should meet several requirements: it should be a low complexity algorithm and the watermark should be embedded and directly detected in the compressed stream to avoid computationally demanding operations.

In general, an image authentication system should satisfy the following criteria:

1) **Sensitivity**: the system must be sensitive to malicious manipulations (e.g., modifying the image *meaning*) as cropping or altering the image in specific areas;
2) **Tolerance**: the system must tolerate some loss of information (originating from lossy compression algorithms) and more generally non-malicious manipulations (e.g., by multimedia providers or fair users);
3) **Localization of altered regions**: the system should be able to precisely locate any malicious alteration made to the image and verify other areas as authentic or corrupted/manipulated;
4) **Reconstruction of altered regions**: the system may need the ability to restore, even partially, altered or destroyed regions in order to allow the user to know the original content of the manipulated areas.

A. Fragile watermarks methods

The main idea underlying these techniques [14] is to insert a specific watermark (independent of the image data) so that any attempt to alter the content of an image will also alter the watermark itself. Therefore, the authentication process consists of locating watermark distortions in order to locate the regions of the image that have been tampered with. The major drawback of these approaches is that it is difficult to distinguish between malicious and non-malicious attacks.

B. Semi-fragile watermarks methods

Semi-fragile watermarks aim to prevent tampering and fraudulent use of modified images. They monitor the integrity of the image content but not its numerical representation. Then, the watermark is designed so that the integrity is proven if the content of the image has not been tampered with, despite some mild processing on the image [15].

C. Block-based watermarks methods

Block-based watermarking techniques [11] consist in dividing the image into blocks of about 64×64 pixels and inserting a "robust" mark into each block. To check the integrity of an image, the authenticator tests the presence or absence of the mark in all blocks. If the mark is present with a high probability in each block, we can affirm that the tested image is authentic.

IV. Attack Techniques

Given the state of contemporary and historical intellectual property use and abuse, for example that of VHS tapes, CDs and DVDs as discussed by Petitcolas [3], the current interest is not only focused in embedding watermarking data, but also in better understanding how the wider and wider consumers community and malicious commercial entities attempt to circumvent, break or totally remove these copyright measures.

In particular, watermark attacks aim to completely strip or prevent the use of a watermark, while preserving the commercial quality of the media [16]. Watermarking attacks can

be classified as “Removal attacks”, “Geometrical attacks”, “Cryptographic attacks” and “Protocol attacks”.

Due to space limits, the following sections review the very basic principles of these attack classifications, except for the cryptographic attacks, which are similar to PKI infrastructure attacks, for example brute-force search for a private key.

A. Removal attacks

Removal attacks are intended to completely remove a watermark from a watermarked image [9]. Three kinds of removal attack are identified: statistical or de-noising, averaging and collusion, and geometrical attacks.

1) Statistical attacks: Statistical attacks treat the watermark as signal noise which can be statistically modeled and removed. Statistical attacks do not need to explicitly model the watermark, its identification and removal may occur as part of another technique such as image filtering (de-noising), remodulation, or the application of a lossy compression algorithm.

Averaging and Collusion attacks work on a set of uniquely watermarked copies of the same image, finding averages across the images which it can be assumed represent the original image with any noise removed as outliers.

B. Geometrical attacks

Geometrical attacks [8], [19], [16], [18], [3] aim to obscure the watermark, making it difficult to detect. This may be achieved by distorting the image and watermark data, degrading the watermark detector’s ability to synchronize with the watermark and resulting in watermark detection failures; essentially the watermark *hidden* in the noise becomes *lost* in the noise. These changes are enough to break a watermark, but in such a way that the human eye or brain cannot see the change in the image.

Geometrical attacks may change global media parameters, such as rotation, aspect ratio and shearing or cropping of the image. Local changes include localized averaging, swapping or removal of pixels, slight color variations, introduction of additional noise, and whatever may introduce visible change to the image, but will confuse a given watermark protocol.

C. Protocol attacks

Voloshynovskiy et al. [16] describes protocol attacks as “attacking the concept of the watermarking application”. That is to say that rather than remove or distort the watermark we may interfere with its intended application, for example by re-watermarking an image with a second owner, and so confusing the media’s original ownership.

1) Mosaic attack: The mosaic attack is of quite general application, and could be tailored to specific watermarking protocols used by particular media vendors. Indeed, as described by Petitcolas [3], this is how the initial work on the attack arose.

A web-crawler, paired with a content distribution mechanism, form an automated online piracy detection system. The web-crawler scans the Internet for images and checks for the distributor’s mechanism which, if detected, can be recorded and checked for licensing issues.

The attack technique is to chop an image into small tiles, which can be imperceptibly rendered by a browser to look the same as a single image. The issue here for the watermark being not only the geometrical attack (cropping) but also that the remaining image may be too small to hide a meaningful watermark inside, such as a single pixel.

Further discussion of issues with online services such as Flash, Java and Javascript mechanisms for image distribution, automated sales, paywalls, etc. are also discussed briefly by Petitcolas. These issues are not related to the stenographical techniques or weaknesses of a given watermark technique, but rather to the protocol of its application, hence they are also sometimes called“Protocol attacks”.

2) Copy Attack: A particularly interesting protocol attack is the copy attack [17]. This pertains to the forgery of watermarks. According to [18], “The goal of the attack it to copy a watermark from stego data to the target data without having any specific knowledge about the watermarking technology” and “the goal of the attack is not to destroy the embedded watermark, but jeopardize the application for which digital watermarks are used” [17]. Therefore, copy attacks are relevant in situations where watermarks (or fingerprints) are used to prove the authenticity or origin of an image rather than to trace copyright ownership. As an analogy consider public/private key signing. If a bank’s SSL certificate could be copy-attacked, then transmissions could be signed without ever needing their private key. This is prevented by hashing in PKI, but media stenography works under fundamentally different constraints.

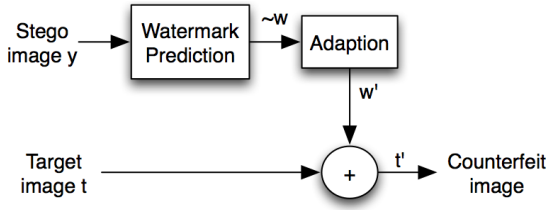

Figure 2. Abstract Copy-attack procedure.

An outline of the Copy attack procedure can be seen in Figure 2. The stego (watermarked) image contains the watermark to be copied. The target image is the image to which a counterfeit watermark will be applied. The predicted watermark is fitted to the target image; its signature should be strong, but it should not produce visible artifacts the new image’s noise is in a different place.

Watermark prediction may involve de-noising techniques, or any other approach to statistically modeling an approximate watermark for adaption and inclusion in the target im-

age. Copy-attack resistance techniques have been proposed as in [12], where video properties such as keyframe distance are used to build the watermark; by coupling the watermark to intrinsic properties of the cover image or video it will become difficult or impossible to copy watermarks directly to a different image.

However, any such scheme producing data-coupled watermarks should pay heed to geometric or re-compression attacks (e.g. resampling a video at a new frame-rate) which will prevent the original watermark from functioning. Which combination of these issues is important relies on the particular application of a watermark; in some cases it may be acceptable for watermarks to be destroyed where the underlying media is somehow perturbed.

V. Conclusions

Digital watermarking applications are becoming widespread in the modern Information and Communications Society (ICS). Although first used in forensics and security systems, the wide uptake of multimedia computers and mobile devices in the consumers market has encouraged a very wide range of practical and novel watermark-based applications to be developed. An exhaustive list of digital watermarking applications is hard to come by, though it is interesting to note this increasing interest and demand from not only consumers and content providers, but other sectors which may or may not traditionally have been interested in. Especially promising are applications related to the copy-protection of printed media. Various companies have projects in developing alternative ways of providing different applications and it is very likely that fully functioning solutions will soon be available.

Several watermarking algorithms exists; in contrast to the spatial-domain-based watermarking, frequency-domain-based techniques can embed more bits of watermark and have proved to be more robust to attacks. On-line application of watermarking for video in the spatial domain becomes cumbersome due to associated high computational complexities involved. Similarly, watermarking in the DCT domain needs preprocessing operations such as inverse entropy coding and inverse quantization.

Therefore, it appears clear that there is not a *best* watermarking technique, but the *optimal* scheme to be employed depends on the medium type, on the application requirements, on the robustness and computational complexity tradeoff, and on the on-the-fly or pre/post-processing operations possibilities.

References

[1] M. Sharkas, D. ElShafie, N. Hamdy, A dual digital-image watermarking technique, Engineering and Technology 5 (2005).

[2] H. Berghel, Watermarking cyberspace, Communications ACM 40 (1997) 19–24.

[3] F. Petitcolas, R. Anderson, M. Kuhn, Attacks on copyright marking systems, in: 2nd workshop on information hiding, Vol. 1525 of LNCS, Springer, 1998, pp. 218–238.

[4] S. Kang, Y. Aoki, Digital image watermarking by fresnel transform and its robustness, in: International Conference on Image Processing (ICIP), Vol. 2, 1999, pp. 221 –225.

[5] B. Mobasseri, M. Marcinak, Watermarking of mpeg-2 video in compressed domain using vlc mapping, in: 7th workshop on Multimedia and security, no. 91-94, 2005.

[6] W. Puech, M. Chaumont, O. Strauss, A reversible data hiding method for encrypted images, in: SPIE Electronic Imaging, Security, Forensics, Steganography, and Watermarking of Multimedia Contents, Vol. 6819, 2008.

[7] J. Wang, R. Healy, J. Timonel, Perceptually transparent audio watermarking of real audio signals based on the cspe algorithm, symposium on Computers and Communications, 2010.

[8] Unzign watermark removal software [online] (1997).

[9] J. Eggers and B. Girod, Quantization effects on digital watermarks, Signal Processing, vol. 81(2), pp. 239263, 2001.

[10] W. Rosenblatt, W. Trippe, S. Mooney, Digital Rights Management: Business and Technology, Hungry Minds, Inc, 2001.

[11] A. Bors, I. Pitas, Image watermarking using block site selection and dct domain constraints, Optics Express 3 (12) (1998).

[12] Y. Ding, X. Zheng, Y. Zhao, G. Liu, A video watermarking algorithm resistant to copy attack, in: 3rd Int. Symposium on Electronic Commerce and Security (ISECS), 2010, 289-292.

[13] K. Kanagawa, K. Hirotsugu, An image digital signature system with zkip for the graph isomorphism, in: International Conference on Image Processing (ICIP), 1996, pp. 247–250.

[14] J. Fridrich, M. Goljan, A. Baldoza, New fragile authentication watermark for images, Int. Conf. on Image Processing, 2000, pp. 446 – 449.

[15] O. Ekici, B. Sankur, Comparative evaluation of semifragile watermarking algorithms, J. Electronic Imaging 13 (1), 2004.

[16] S. Voloshynovskiy, S. Pereira, V. Iquise, Attack modelling: Towards a second generation watermarking benchmark, Signal Processing 81 (6) (2001) 1177–1214.

[17] M. Kutter, S. Voloshynovskiy, Watermark copy attack, in: SPIE, Vol. 3971, 2000.

[18] A. Soni, A. Goel, O. Sahu, P. Soni, A copy attack on robust digital watermarking in multi domain for the stego images, International Journal of Research and Reviews in Computer Science 1 (2) (2010) 47–49.

[19] stirmark benchmark 4.0 [online] (1997).

[20] F. Petitcolas, M. Kutter, A fair benchmark for image watermarking systems, in: T. I. S. for Optical Engineering (Ed.), Electronic Imaging, Vol. 3657, 1999.

[21] P. Blythe, J. Fridrich, Secure digital camera, in: Digital Forensic Research Workshop (DFRWS), 2004, pp. 17–19.