INTRODUCTION TO DIGITAL WATERMARKS AND CLASSIFICATION OF ATTACKS

ESE558 DIGITAL IMAGE PROCESSING
Instructor: Dr. Murali Subbarao

Submitted by:
Abhishek Goswami(105209898),
Graduate Student,
Department of Electrical and Computer Engineering,
Stony Brook .

TABLE OF CONTENTS

LIST OF FIGURES

INTRODUCTION TO DIGITAL WATERMARKING

The Internet is an excellent sales and distribution channel for digital assets, but copyright compliance and content management can be a challenge. These days, digital images can be used everywhere – with or without consent. Images that are leaked or misused can hurt marketing efforts, brand image and, ultimately, sales. With one click, your digital assets can be detached from your copyright information, so guarding brand and intellectual property assets is essential.
Watermarking solutions let you add an extra layer of protection to your digital images.

DIGITAL WATERMARKING OF DIGITAL IMAGES ENABLES:

Copyright Protection

Embed copyright, owner ID and other digital information into digital images, telling who owns it and how it can be used.

Brand Image Tracking

Track where and how your brand assets are being used on the Internet, for simplified management of online promotions and channel copyright compliance. If you're a photographer, a web publisher, or an image distributor or if you maintain a collection of images for a corporation or museum, you should consider digitally watermarking your images any time they are released externally (licensing to a magazine, posting on the World Wide Web, adding to a stock image collection, etc.). Digital watermarking gives you the security of knowing that no matter how or where your images appear online, they carry your notice of ownership and a simple path to contact you through the registry.

WHAT IS DIGITAL WATERMARK?

A digital watermark is best described by comparing it to a traditional paper watermark. Traditional watermarks are added to some types of paper to offer proof of authenticity. They are imperceptible, except when the paper is held up to a light for inspection. Similarly, digital watermarks are added to digital images in a way that can be seen by a computer but is imperceptible to the human eye. A digital watermark carries a message containing information about the creator or distributor of the image, or even about the image itself. A digital watermark is used to communicate copyright information about an image in order to reduce copyright infringement. A person opening a digitally watermarked image in a image-editing application or our Internet- or Windows-Explorer reader receives notification through a copyright symbol ((c)) that the image contains copyright and ownership information. The digital watermark can provide a link to complete contact details for the copyright holder or image distributor, making it easy for the viewer to license the image, license another one like it, or commission new work. Digital watermarks are imperceptible to the human eye, yet provide images with a durable, persistent identity. To help hide the digital watermark, varies the digital watermark energy within the image so that it remains imperceptible in both flat and detailed areas. The digital watermark is robust, surviving many typical image edits and file format conversions.

HOW THE DIGITAL WATERMARKING SYSTEM WORKS?

When combined, digital watermarking products and services form a complete copyright communication and image tracking system for digital images.

This system provides the tools and capabilities to:
• Embed digital watermarks into images
• Detect and read digital watermarks
• Link to complete contact details or a web site for the image creator or distributor (for inquiring about usage rights, licensing, etc.)
• Track instances of digitally watermarked images on the web.

With the growth of numerical technologies, it became extremely easy to reproduce a data without any damage. For instance, any image taken on the Internet can be saved for a personal usage and then be written on a CD-Rom or on another web page.

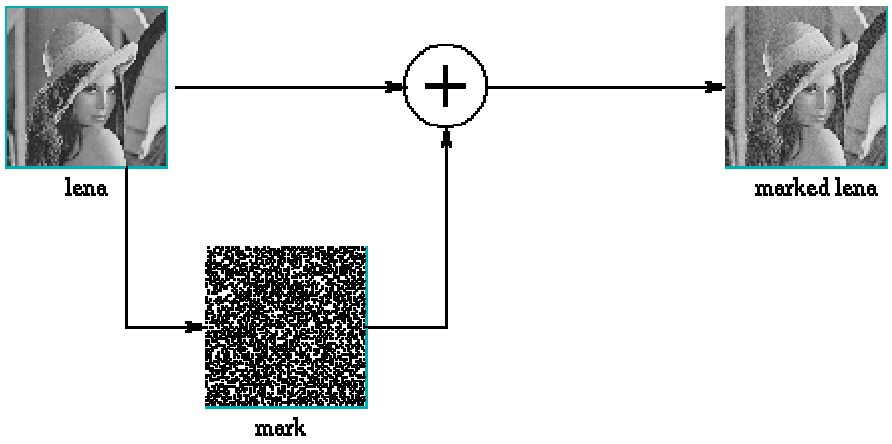


Fig1: Example of Independent Watermarking

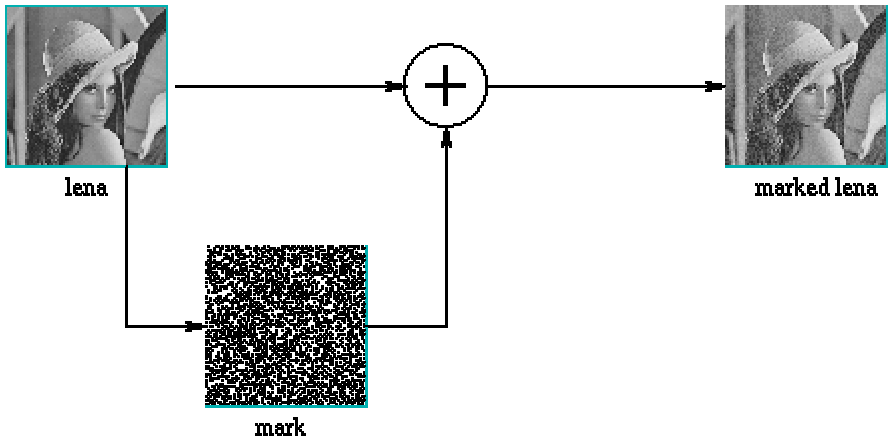


Fig 2: Example of Independent Watermarking

To summarize it is mainly composed of three steps :

1. Mark Computation: the mark itself must have some important properties;
2. Mark Positioning: it is important to embed the mark in an image in an effective way, in order to insure some of the basic mark properties;
3. Mark Recovery: the main problem of this step is to avoid "false positives" (recognize a wrong signal as a mark).

OTHER FACTORS AFFECTING DIGITAL WATERMARK STRENGTH

Along with the intensity setting that you choose when embedding a watermark, the strength of a digital watermark is also affected by the following factors:

• Image variations/randomness: As discussed earlier in the "Image variations/randomness" section, the successful embedding of a digital watermark is dependent on the variation and randomness present in the pixels making up the image. For example, if you are working with an image that contains more flat color regions than detailed areas, you may want to choose a higher digital watermark strength so that the watermark will overcome the limitations of the specific image. This may result in a more visible digital watermark, but in some situations that is an acceptable trade-off, as mentioned above.

• Image size: As far as possible the Watermarks should be immune to Image size.

• Compression: Saving the watermarked image in a compressed format may affect the durability of the digital watermark.

The following factors will influence the impact that lossy compression has on digital watermark survival:
• Level of image compression: Lossy compression degrades the image to some extent, depending upon the quality setting chosen when saving in compressed format; most digital watermarks will survive as long as a moderate level of compression is used (see below for more detail).

• Visibility/durability setting used when embedding a digital watermark: The higher the intensity setting, the better the chances the digital watermark will survive compression. Again, a higher-intensity digital watermark provides more data-to-survive compression. Since the visual quality of compressed images is often somewhat compromised anyway, generally a higher watermark intensity setting yields quite acceptable results.

• Image size: The greater the number of pixels in the image, the more the digital watermark can be repeated throughout it; the recommended minimum size for an image that will be compressed is 256 x 256 pixels. The larger the image, the better the digital watermark will survive compression.

• Randomness of image data: As discussed in the earlier section "Image variations/randomness," the more randomness and/or color variation in an image, the better; a flat color space with little gradation may not survive well, while an image with more

detail and contrast will fare better. Since a digital watermark is applied more strongly within areas of high contrast or variation, an image that contains more contrast and/or variation than others will contain more digital watermark data and thus stand a better chance of surviving compression.

The watermark can be regarded as an additive signal w, which contains the encoded and modulated watermark message b under constraints on the introduced perceptible distortions given by a mask *M* so that

$$J' = x + w(M).$$

Note that w need not be independent from the original data x. The simplest approach *to* achieve a perceptually indistinguishable watermarked and original signal is to keep the power of the watermark signal very low. Using sophisticated psycho-acoustic or psycho-visual models, more appropriate masks *M* can be applied to enhance the robustness of the watermarking scheme. Commonly used embedding techniques can be classified into additive [2], multiplicative[2], and quantization-based schemes [3, 4]. In additive schemes, there are usually very weak dependencies between w and *x* (e.g., introduced by choosing *w* dependent on a data-dependent perceptual mask *M).* In multiplicative schemes, samples of the original data are multiplied by an independent signal v so that $w = xv - x$. Here, w and x are of course dependent on each other. Strong local dependencies between the realizations of w and x exist in quantization-based watermarking schemes. However, these dependencies are such that statistically x and w appear (almost) independent. To be precise, we have to distinguish between the watermark signal *w,* which is the actual signal added to the original data, and the watermark message or information *b* that is conveyed by the watermark signal. Usually the meaning is clear from context. Coding schemes can be used to achieve reliable watermark communication. In some cases only one bit needs to be communicated (on-off signaling), while in other cases a sequence of M-ary watermark symbols is transmitted.

In most watermarking applications, the marked data is likely to be processed in some way before it reaches the watermark receiver. The processing could be lossy compression, signal enhancement, etc . An embedded watermark may unintentionally or inadvertently be impaired by such processing. Other types of processing may be applied with the explicit goal of hindering watermark reception. In watermarking terminology, an "attack" is any processing that may impair detection of the watermark or communication of the information conveyed by the watermark.

Broadly it can be classified as –

- Intentional Attacks
- Non-Intentional Attacks

The processed watermarked data is then called "attacked data". An important aspect of any Watermarking scheme is its robustness against attacks. The notion of robustness is intuitively clear: A watermark is robust if it cannot be impaired without also rendering the attacked data useless. Watermark impairment can be measured by criteria such as miss probability, probability of bit error, or channel capacity. For multimedia, the usefulness of the attacked data can be gauged by considering its perceptual quality or distortion. Hence,

robustness can be evaluated by simultaneously considering watermark impairment and the distortion of the attacked data. An attack succeeds in defeating a watermarking scheme if it impairs the watermark beyond acceptable limits while maintaining the perceptual quality of the attacked data.

STATE-OF-THE-ART WATERMARKING ATTACKS

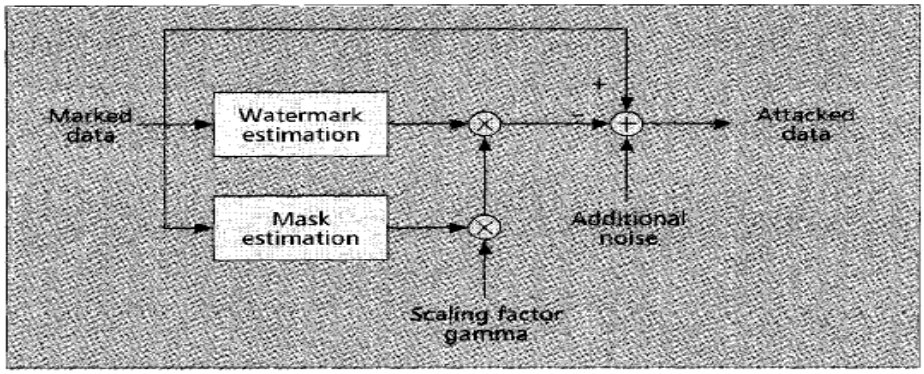
One categorization of the wide class of existing attacks contains four classes of Attacks:

- Removal Attacks
- Geometric Attacks
- Cryptographic Attacks
- Protocol Attacks.

REMOVAL ATTACKS

Removal Attacks aim at the complete removal of the watermark information from the watermarked data without cracking the security of the watermarking algorithm (e.g., without the key used for watermark embedding). That is, no processing, even prohibitively complex, can recover the watermark information from the attacked data. This category includes denoising, quantization (e.g., for compression), remodulation, and collusion attacks. Not all of these methods always come close *to* their goal of complete watermark removal, but they may nevertheless damage the watermark information significantly. Sophisticated removal attacks try to optimize operations like denoising or quantization to impair the embedded watermark as much as possible while keeping the quality of the attacked document high enough. Usually, statistical models for the watermark and the original data are exploited within the optimization process.

Collusion attacks are applicable when many copies of a given data set, each signed with a key



■ **Figure 1.** *A perceptual remodulation attack.*

or different watermark, can be obtained by an attacker or a group of attackers. In such a case, a successful attack can be achieved by averaging all copies or taking only small parts from each different copy.

GEOMETRIC ATTACKS

In contrast to removal attacks, *geometric attacks* do not actually remove the embedded watermark itself, but intend to distort the watermark detector synchronization with the embedded information. The detector could recover the embedded watermark information when perfect synchronization is regained. However, the complexity of the required synchronization process might be too great to be practical. However, most recent watermarking methods survive these attacks due to the use of special synchronization techniques. Robustness to global geometric distortions often relies on the use of either a transform-invariant domain (Fourier-Melline) or an additional template, or specially designed periodic watermarks whose auto-covariance function (ACF) allows estimation of the geometric distortions. However, as discussed below, the attacker can design dedicated attacks exploiting knowledge of the synchronization scheme.

Robustness to global affine transformations is more or less a solved issue. Therefore, pixels are locally shifted, scaled, and rotated without significant visual distortion. However, it is worth noting that some recent methods are able to resist this attack.

CRYPTOGRAPHIC ATTACKS

Cryptographic attacks aim at cracking the security methods in watermarking schemes and thus finding a way to remove the embedded watermark information or to embed misleading watermarks. One such technique is brute-force search for the embedded secret information. Another attack in this category is the so-called Oracle attack, which can be used to create a non-watermarked signal when a watermark detector device is available. Practically, application of these attacks is restricted due to their high computational complexity.

PROTOCOL ATTACKS

Protocol attacks aim at attacking the entire concept of the watermarking application. One type of protocol attack is based on the concept of invertible watermarks[7]. The idea behind inversion is that the attacker subtracts his own watermark from the watermarked data and claims to be the owner of the watermarked data. This can create ambiguity with respect to the true ownership of the data. It has been shown that for copyright protection applications, watermarks need to be noninvertible. The requirement of non-invertibility of the watermarking technology implies that it should not be possible to extract a watermark from a non-watermarked document.

A solution to this problem might be to make watermarks signal-dependent by using one-way functions. Another protocol attack is the copy attack. In this case, the goal is not to destroy the watermark or impair its detection, but to estimate a watermark from watermarked data and copy it to some other data, called target data [8]. The estimated watermark is adapted to the local features of the target data to satisfy its imperceptibility. The copy attack is applicable when a valid watermark in the target data can be produced with neither algorithmic knowledge of the watermarking technology nor knowledge of the watermarking key. Again, signal-dependent watermarks might be resistant to the copy attack.

ESTIMATION- BASED ATTACKS

Here, we consider attacks that take into account the knowledge of watermarking technology and exploit statistics of the original data and watermark signal [5, 8-32]. In addition, we emphasize that for the design of attacks against watermarking schemes, the distortion of the attacked document and the success of watermark impairment has to he considered. Within the scope of these attacks, we present the concept of estimation-based attacks. This concept is based on the assumption that the original data or the watermark can be estimated - at least partially - from the watermarked data using some prior knowledge of the signals' statistics. Note that estimation does not require any knowledge of the key used for watermark embedding. Furthermore, knowledge of the embedding rule is not required, but the attack can be more successful with it. Depending on the final purpose of the attack, the attacker can obtain an estimate of the original data or of the watermark based on some stochastic criteria such as maximum likelihood (ML), maximum a posteriori probability (MAP), or minimum mean square error (MMSE). We do not focus here on the particularities of the above estimation but rather concentrate on different ways to exploit the obtained estimates to impair the embedded watermark. Depending on the way the estimate is used, we can classify estimation-based attacks as removal, protocol, or desynchronization attacks.

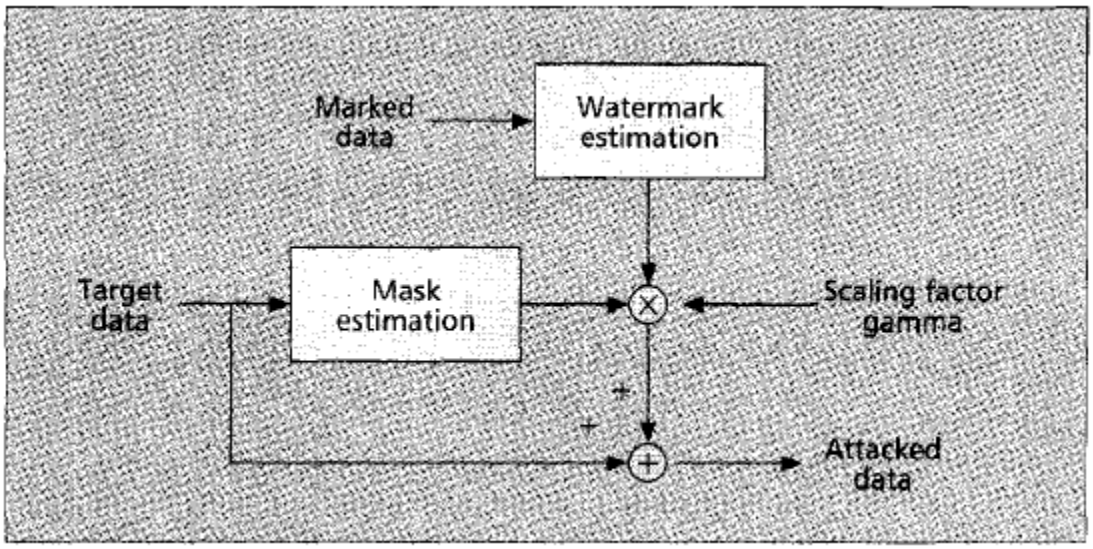
REMODULATION ATTACKS

Remodulation attacks aim at modification of the watermark using modulation opposite to that used for watermark embedding. Assuming the estimated watermark is correlated with the actual watermark, meaning a good estimate could be obtained, the estimated watermark can be subtracted from the watermarked data. Subtracting a very inaccurate estimate of the watermark might decrease the document quality without affecting the watermark too much. On the other hand, correlation-based detection can be defeated by subtracting an amplified version of the estimated watermark. For this reason, we introduced a gain factor y *2* 1, which gives us the possibility to trade off the distortion of the attacked document vs. the success of the attack. There are four basic variations of the remodulation attack. First, when y = 1, the attack yields the MMSE estimate of the original and reduces to the denoising attack.

Second, for y > 1, the quality of the attacked document might be reduced, but correlation based detection might be defeated more successfully. The attack can even drive the correlation to zero so that the detector incorrectly decides that the watermark is not present in the attacked data. Third, when using a more sophisticated distortion measure than simple MSE, a better compromise between success of the attack and introduced distortion can be obtained by weighting the re-modulated watermark by a perceptual mask. Fourth, the attacker can not only subtract the weighted, estimated watermark, but also add outliers to obtain a non-Gaussian noise distribution, which decreases the performance of correlation-based detection. Moreover, exploiting features of the human perceptual system, the attacker can efficiently embed a large amount of outliers in perceptually less significant parts of the data. For image data, this approach has been demonstrated to be successful in [9]. We rcfer to this attack as Perceptual Remodulation.

COPY ATTACK

The estimated watermark can be exploited to implement a copy attack, as already described. Of course, the copied watermark has to be adapted to the target data to keep the quality of the falsely watermarked target data high enough. There are many practical ways to adapt the watermark to the target data based on perceptual models. For images, contrast sensitivity and texture masking phenomena of the HVS can be exploited. The estimation-based copy attack is most successful when the same perceptual model is used as in the original watermarking algorithm. Note that the copy attack in its described version is mainly applicable to additive watermarking schemes. In the case of quantization-based watermarking schemes, even a perfectly estimated watermark signal w cannot be copied since it is highly unlikely that the copied signal w is a valid watermark in the target signal (Fig. 2).



■ **Figure 2.** *A copy attack.*

ATTACKS DEPENDENT ON LOCAL SIGNAL STATISTICS

Different estimation-based attacks can successfully be combined depending on local signal statistics. Here we focus on image watermarking. The attacker is motivated to reduce the maximum rate of reliable communication also by exploiting the HVS and the possibility to remove the watermark based on different models of the image. The watermark can easily be predicted and removed from flat image areas rather than from areas of edges and textures. The stochastic model for the edges and textures is nonstationary and relatively complicated to use for accurate image estimation. Therefore, the attacker will try as much as possible to utilize the advantages of denoising and remove the watermark from flat areas without visual distortions and even enhancing the PSNR. In contrast, the attacker will use remodulation with increased strength in the edges and texture areas, which again will he masked by the HVS. At the same time, the attacker can use the NVF to automatically determine the flat regions, edges, and textures. This attack is schematically shown in Fig. 5b. The practical power of this attack was first demonstrated in [Y], and a further theoretical analysis can be found in [5].

OPTIMIZED ATTACKS

So far, we have developed different attacks based on watermark estimation and discussed how an embedder can make watermark estimation as difficult as possible. However, the embedder has another chance to react to estimation based attacks, especially to the remodulation attack. The detector can first estimate the remodulated watermark and try to invcrt the remodulation, and thus retain reliable watermark detection. For instance, when watermark estimation is based on Wiener filtering, the detector can apply inverse Wiener filtering. This is of course not wanted by the attacker. Thus, he/she has to add noise to the attacked data, which would be amplified by the inverse Wiener filtering and thus impair watermark detection.

Of course, the additive noise further degrades the attacked signal. Now, we have the situation where the attacker has to find a good combination of using the estimated watermark and the noise to be added. The embedder no longer simply tries to make watermark estimation as hard as possible, however; he .also has to get a power advantage over the additive noise an attacker might introduce. This problem can be formulated in an even more general way: the attacker attempts to minimize the watermark capacity under a constraint on the distortions introduced by the attack. The embedder attempts to maximize the watermark capacity under *a* constraint on the embedding distortions. This situation can be regarded as a game between the attacker and embedder. To solve this problem, we followed a game-theoretic approach, assuming that the embedder and attacker know each other's behavior .

CONCLUSIONS

The article presents an introduction to digital image watermarking meant for students who are taking an introductory course on Image Processing. It gives a brief overview of the necessity of watermarking the images, steps followed to watermark and recover the image, factors affecting the recovery etc. It also elucidates the difference between encryption and watermarking. Further, attacks on digital watermarking systems are investigated in this article. First, a categorization of different attacks is given, and popular attacks are briefly described. We point out that attacks on digital watermarks must consider both watermark survival and the distortion of the attacked document. Early attacks do not exploit as much knowledge of the watermarking scheme as possible; also, they do not consider the distortion of the attacked document. Since attacks can be improved by using knowledge of the watermarking scheme and the signal statistics, we take a look at the new set of attacks called estimation-based attacks. The general idea is to estimate the watermark and exploit it to trick the detector. It is shown that this approach is related to denoising, but can be extended to a variety of different attack methods. Finally, we explain how considering the watermarking and attacking problem as a game between embedder and attacker can be exploited to find the watermark capacity, when facing an optimized attack with a constrained attack distortion. The theoretical analysis of watermark attacks gives many insight into the watermarking problem, and enables us to show fundamental limits of this technology.

BIBLIOGRAPHY

[1] M. Kutter and F. Petitcolas, "A Fair Benchmark for Image Watermarking Systems," Electronic Imaging *'99:* Security and Watermarking *of* Multimedia Content, SPIE Proc., vol. 3657, San Jose, CA, Jan. 1999.
[2] I. Cox et al., "Secure Spread Spectrum Watermarking for Multimedia," */E€€* Trans. Image. Proc., vol. 6, no 12, Dec. 1997, pp. 1673-87.
[3] B. Chen and G. W. Wornell, "Dither Modulation: A New Approach to Digital Watermarking and Information Embedding," Security and Watermarking *of* Multimedia Contents, Proc SPIE. vol 3657, San Jose, CA, Jan. 1999.
[4] J. J. Eggers, J.K. Su, and B. Girod, "A Blind Watermarking Scheme Based on Structured Codebooks," Colloq: Secure Images and Image Authentication. London, UK, Apr. 2000.
[5] S. Voloshynovskiy et al.. "Attack Modeling: Towards a Second Generation Watermarking Benchmark," Sig. Processing. Special Issue on Information Theoretic Issues in Digital Watermarking, 2001, vol. 81, no. 6, pp. 1177-214.
[6] Sviatolsav Voloshynovskiy, Shelby Pereira, and Thierry Pun, University of Geneva Joachim 1. Eggers and Jonathan K. Su, University of Erlangen - Nuremberg,"Attacks on Digital Watermarks: Estimation-Based Attacks, and Benchmarks", 0163-6804/01/0 2001 IEEE Communications Magazine August ZOO1

WEBSITE LINKS

1]http://academic.mu.edu/phys/matthysd/web226/L0205.htm
2]http://www.profc.udec.cl/~gabriel/tutoriales/rsnote/cp10/cp10-9.htm
3]http://www.cg.tuwien.ac.at/~theussl/DA/node36.html
4]http://www.cs.umsl.edu/~sanjiv/classes/cs5420/lectures/spatial.pdf