PDF Digital Watermarking: Grayscale Photocopy Detection of Printed Documents

Norhaziah Md. Salleh[a], Soong Hoong Cheng[b]

[a]*Faculty of Information and Communications Technology*
Universiti Teknikal Malaysia Melaka, Ayer Keroh, Melaka
Tel : 06- 233 2494, Fax : 06-2332508
E-mail : haziah@utem.edu.my

[b]*Faculty of Information and Communications Technology*
Universiti Teknikal Malaysia Melaka, Ayer Keroh, Melaka
Tel : 017-5627755, Fax :06-2332508
E-mail : hoongcheng@hotmail.com

ABSTRACT

Digital watermarking is the technique used to provide additional and useful evidences for many application fields especially in copyright infringement detection. The focus of this paper is on Portable Document Format (PDF) watermarking and the printed watermarked copies. Digital watermarking is indeed suitable to detect unauthorized copies from any authentic sources. Colour theory and colour properties are studied on how to prevent yellow-watermarked stamps being photocopied in grayscale, particularly by using luminance concept on the documents. Please note that the technique applies only for grayscale photocopies, as the detection for coloured copies requires disparity techniques for the coloured detection.

Keywords:
Digital Watermarking, Grayscale, Document Management System, Luminance, Relative Luminance, PDF Watermarking

1.0 INTRODUCTION

According to Ross et al. (1975), the purpose of hiding information is to make inaccessible certain details that should not affect other parts of a system. As such, information hiding techniques have recently been given much attention by researchers and industries. Thus, the interest in information hiding was triggered by concerns over copyright issues as audio, video, documents, and other works are now available in digital formats making it easier to make unauthorized copies. In consequence, watermarking is one of the techniques for information hiding. Apart from that, digital watermarking is defined by Kutter (2000-2003) as imperceptible insertion of information into multimedia data. At most (Jim Meehan et al., 2005), Portable Document Format (PDF) digital documents are popular nowadays due to the *de facto* standard for electronic exchanges of documents and are now, as the standard in the industries for intermediary representation of printed material in electronic prepress systems for conventional printing applications.

Moreover, it can be observed that many types of documents are used in a manufacturing organization. Some of them include assembly instructions, engineering memos and drawings, bills of materials, procedures, diagrams, and work instructions. Most of these documents are considered as controlled documents and as such only one version of the same document can be held by employees at any one time. As a result, tracking of these document versions is difficult and prone to errors when there are many documents circulating in the organization. Hence, the distribution of documents must be properly controlled and managed to ensure only valid documents are in circulation and they are not unauthorized copies.

This raises the issue of how to detect whether the printed copies are authentic or otherwise. In the industries where only grayscale photocopiers are commonly available, this research focuses mainly on detecting unauthorized grayscale photocopies by preventing yellow-watermarked stamps being transferred to other printed copies. As an added security measure, the owner's details are imprinted on the original hardcopies. A system was developed to control documents in a manufacturing environment to track the authentic copies. Employees in an organization may borrow documents for reference and each borrowed document will be printed in yellow colour with the word "Confidential" across each page of the document and the borrower's details such as name, staff number, and date printed and time will be stamped onto the document. These measures were made to prevent the document from being copied and distributed illegally within the organization as the photocopiers are usually not colour copiers. In grayscale copying the yellow word cannot be copied. Even though the detection is unable and impossible to prevent photocopying from the printed materials, aforementioned steps are enough to halt illegally duplicated copies. Apart from that, the user information watermarked on the genuine document copies enables the system to expose the masqueraders claiming to be the valid owners of the original copies as well as the duplicated copies.

2.0 RELATED RESEARCH

As mentioned before, the focal concentration is in PDF digital watermarking. What is exactly digital water marking? According to van Schyndel et al. (1994) and Cox et al. (1996), digital watermarking is used as the last resort and as the "last line of defenses" as the failure safeguard of the encryption or copy protection occurred and thus the illegal copies are recognized as to against distribution of valuable digital media. Furthermore, Su et al. (1999) stated that a digital watermarking is embedded directly into the system. As for the exemplar, information about copyrights, ownership, timestamps, and the legal recipient possibly will be embedded. However, the implanted digital watermarking by itself is unable to avoid illegally copying, modification and re-distribution of the genuine documents. Nevertheless, watermarking is efficient in tracking and tracing to the rightful owners of detecting unauthorized usage of documents. Thus, punishable actions can be taken to the illegal users provided the watermark can be retrieved and detected from the documents whether it is invisibly encrypted or visibly stamped as evidences.

For that motive, the watermark should be excessively robust to prevent modification and altering of the watermark in the original copies but not necessary for the duplicated copies. Thus, robustness generally should not be avoided if possible and it is well defined by Kutter & Petitcolas (1999) for a digital watermark if it has the properties as listed in the following Table 1:

Table 1: *Types of digital watermarking robustness.*

General Robustness	General Robustness Subtypes
JPEG compression	-
Geometric transformations	Horizontal flip, Rotation, Cropping, Scaling, Deletion of lines or columns, Generalized geometrical transformations, Random geometric distortions (StirMark), Geometric distortions with JPEG
Enhancement techniques	Low pass filtering, Sharpening, Histogram modification, Gamma correction, Color quantization, Restoration
Noise addition	-
Printing-scanning	-
Statistical averaging and collusion	-
Over-marking	-
Oracle attack	-

Imperceptible watermark is the direct results of the robustness which occurred. Imperceptible watermark provides the property of being indistinguishable between copied watermarks and the original watermarks. Generally, an effective watermark shows several properties besides specifically being robust. The properties are robustness, imperceptibility or a low degree of obtrusiveness, security, fast embedding and/or retrieval, no reference to original document, multiple watermarks, and not ambiguity (Su et al., 1999). Apart from that, watermarking usually has a set watermarking scheme that has to be done regardless of any type of digital watermarking. The watermarking scheme is as in the equation (1) as given by Dittmann et al. (2006) below:

$$Sc = (E, D, R, M, Pe, Pd, Pr) \qquad (1)$$

Sc represents the instance of watermarking scheme, E is the embedding method, D is the detecting function, R is the retrieval function, M is the messages, Pe is watermarking parameters, Pd defines the detection parameters, and finally Pr is the retrieval parameters. According to Wolfgang et al. (1998), perceptually based watermarks consist of three principles that are robustness, capacity, and transparency. Consequently, digital watermarking can indeed be classified as robustness, capacity, perceptibility/transparency and lastly the embedding methods. Chen & Wornell (2001) states that the embedding methods can be classified into three categories are such as spread-spectrum, quantization, and amplitude modulation. In spite of using only perceptible watermark imprinted on the document, illegal grayscale photocopy is easily identified as the yellow-watermarked stamps are impossible to be grayscale photocopied. Hence, what is actually grayscale and how is it mapped from the coloured sources with the reference of black and white copies? Based on the definition from Johnson) (2006), as in photography and computing, the intensity of the information is represented by each pixel that is a single sample from a grayscale or greyscale digital image. On the other hand, this type of grayscale image, which is also called as black-and-white, is composed exclusively of shades of gray, varying from black, at the weakest intensity to white, at the strongest. As for this, what are colours then before being converted into grayscale? Colours as defines by Bohren & Clothiaux (2006) are the spectrums of light (distribution of light energy versus wavelength) that can be perceived visually by the eyes while the light receptors interpreting the visual perceptual properties gives the colours for example red, blue yellow and others. In addition, colours have several properties (Table 2) that defines what colour is. Moreover, hue defines colour the most.

Table 2: *Colour properties.*

Colour Properties	Description
Hue	The intensity of the colours.
Chroma / Saturation	Direction of the colours from white.
Shade	Deeper colours from additive black.
Tint	Lighter Colours from additive white.
Value	The brightness of the colours.

From the Munsell system (Kuehni, 2002), the relation between the colours properties are easier to be comprehended as. Going back to the basics, there are indeed two types of colour models available to produce arrays of colours. Therefore, (Hirsch, 2004) additive colour model is the addition of blue, red and green lights to produce secondary colours that are cyan, magenta, yellow and white (all colours combined). As for the RGB colour model with primary colours of red, green and blue, (Poynton, 2003) the RGB colour model is an additive colour model that is similarly producing arrays of colours by mixing the primary red, green and blue lights. On the other hand, the colours can be subtracted to get back the primary colours from the secondary colours of lights. Thus, (Berns, 2000) subtractive colour model is the colours from the mixture of natural colorants that absorb several wavelengths of lights to produce new colours reflecting on it. Usually, subtractive colours are the secondary colours such as cyan, magenta, yellow and black (all colours subtracted) to the primary colours (red, green, blue). Hence, (Gatter, 2005), CYMK colour model is the same as the subtractive colour model and strictly used in printing applications. Welsh et al. (2002) states that grayscale may be obtained from the luminance of the colours in reference to black and white. Hence, colorimetry (Hirakawa & Parks, 2005) is the science of measuring color as well as can be translated into the value of luminance to give a new grayscale colour as in equation 2 (luminosity function).

$$F = 683.002\ \mathrm{lm/W} \cdot \int_0^\infty \overline{y}(\lambda) J(\lambda) d\lambda \tag{2}$$

However, (Stokes et al., 1996) by using sRGB colour space it is much easier to denote the relative luminance as compared to luminosity function. Yet, the relative luminance does reflect the luminosity function from the equation (2). From the colour space as denoted in XYZ, the linear Y is derived as the luminance from the colorimetric measurement as in equation (3). Consequently, as the sRGB colour space utilizing ITU-R BT.709 primaries, the linear RGB components certainly can calculate the relative luminance with the RGB relative intensities of each colours.

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.4124 & 0.3576 & 0.1805 \\ 0.2126 & 0.7152 & 0.0722 \\ 0.0193 & 0.1192 & 0.9505 \end{bmatrix} \begin{bmatrix} R_{sRGB} \\ G_{sRGB} \\ B_{sRGB} \end{bmatrix} \tag{3}$$

Where, Y = 0.2126 R + 0.7152 G + 0.0722 B

3.0 METHODS

The following subsections are the approaches of the PDF digital watermarking that entail the process flow of the system **(system architecture)**, programming techniques **(PDF stamping technique)**, theoretical watermarking technique **(digital watermarking technique and luminance theory)**, and the prevention of document modification **(strategic watermark location)**. Nevertheless, the focus of the PDF watermarking approach is focussed solely on the hardcopies and the flow of the system instead of the digitally imperceptible watermarking as the watermarked softcopies are deleted as soon the documents are sent to the printers. In spite of that, it is essential to learn in depth the digital watermarking technique for the softcopies as it may be useful for the current research as well as for the future research.

3.1 Proposed system: System Framework/Architecture

By using the system process, certain system flows could increase the watermarking robustness as well as to decrease the chances of perceptibility to the imprinted watermark. Accordingly, the system requires certain level of permission loops before acquiring the watermarked printed documents. For this reason, the digital documents are sent directly to the server for the printing process and the clients are denied access to the softcopy, thus protecting the digital watermarks. As a final point, database records track the digital documents and thus the authenticity can be matched for the detections of the printed watermarked documents. Figure 1 illustrates the process flow of the proposed system known as Document Control System (DocCon).

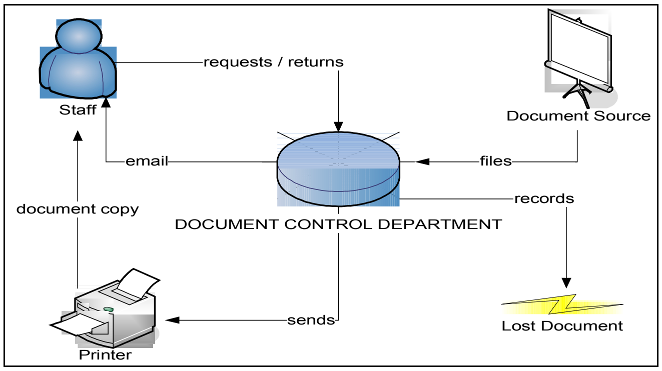


Figure 1: Process flow of proposed system

3.2 PDF Watermarking: Technique

There are two subsections of the PDF watermarking techniques considered as part of the approach. In order to insert a watermark in a document, it is crucial to have an understanding of what and how it is being inserted. The watermarking parameters are studied theoretically from related researches and hence robustness will be the topic of discussion in the subsequent subsection. Understanding of the parameters the programming techniques (appending digital watermark to PDF files) are equally important so as to avoid deviation of the expected results at the final stage of the research.

3.1.1 Digital Watermarking Technique: Embedding Method

According to Wolfgang et al. (1998), the digital watermarking techniques can be classified into several categories as listed in Table 3. As mentioned before, the focus of the PDF stamping is more to the hardcopies instead of the softcopies as the softcopies are deleted after being printed for security reasons and thus only the robustness principal is discussed in this section. Subsequently, the watermark intended for the document from robustness is considered fragile as it is only for detection purposes. From the equation (4) (Dittmann et al., 2006), when another illegally grayscale is photocopied, the fragile watermarks are marked as 0 bit of representation of the watermark as no successful detection is available. As a result, we can conclude when the original stamps are not on the other grayscale copies it can be assumed the copies are unauthorized copies from the original source. As for the other technique, it will not be discussed in detail but it is essential to know the theory of the digital watermarking techniques to enhance or to comprehend better the implementation in stamping the watermarks to the PDF files.

Table 3: *Digital watermarking techniques.*

Techniques	Sub-Techniques
Robustness	-
Capacity	-
Perceptibility	-
Embedding Method	Spread-spectrum, Quantization, Amplitude modulation.

$$\det_D(\Omega^*, S_{EA}) = \begin{cases} 0, \text{no successful detection (negative)}, \\ 1, \text{positive successful detection (positive)}. \end{cases} \quad (4)$$

3.1.2 PDF Stamping Technique: Appending Digital Watermark

According to Fitzgerald (2004), there are steps that must be followed to append the watermark efficiently as illustrated in Figure 2. An extra page must be created prior to be spawned or not spawned to the entire document. As soon as the insertion is completed, the original page and template are deleted to avoid addition of unnecessary pages. A figure 3 shows those steps are applicable to any programming language, as long as the steps are followed. However, even though the steps are similar for each programming technique, the syntax and sometimes the programming methods may vary.

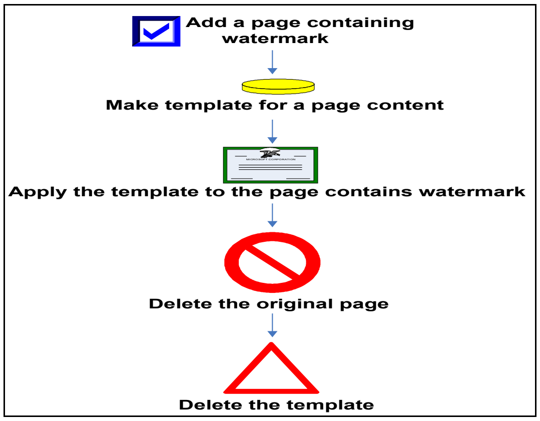
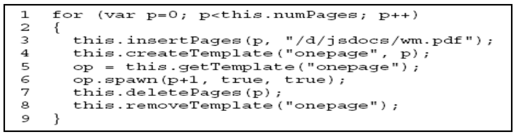

Add a page containing watermark
Make template for a page content
Apply the template to the page contains watermark
Delete the original page
Delete the template

Figure 2: Steps of Appending the Watermarks

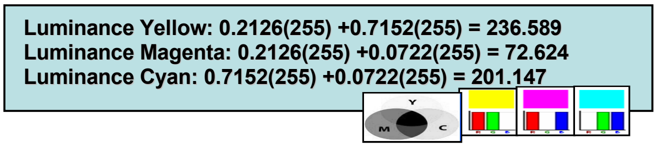


```
for (var p=0; p<this.numPages; p++)
{
  this.insertPages(p, "/d/jsdocs/wm.pdf");
  this.createTemplate("onepage", p);
  op = this.getTemplate("onepage");
  op.spawn(p+1, true, true);
  this.deletePages(p);
  this.removeTemplate("onepage");
}
```

Figure 3: *Fragment code of appending the watermarks (JavaScript)*

3.3 Luminance Theory: Yellow Watermark Approach

Photometry (Ohno, 1999) deals with the measurement of visible light as perceived by human eyes. The human eye can only see light in the visible spectrum and has different sensitivities to light of different wavelengths within the spectrum. When adapted for bright conditions (photopic vision), the eye is most sensitive to greenish-yellow light at 555 nm. Hence, the green constant will be the highest, using the linear function of relative luminance: Y = 0.2126 R + 0.7152 G + 0.0722 B. Given the yellow intensity is (255, 0, and 0), magenta intensity is (255, 0, and 255) and cyan intensity is (0, 255, and 255) from the RGB value, then map the luminance intensity to the scale of grayscale where the whiter colour will consist of higher luminance since white has the maximum relative luminance as in Figure 4. As a result, yellow will be as near to white because of the higher luminosity, compared to the grayscale generated CMYK colour model from Adobe Photoshop CS3 as in Figure 5. From the results, it is easily detected that the document is grayscale photocopied from the original coloured sources. Therefore, this particular detection is appropriate in the working environment where black and white photocopier is commonly available.



Luminance Yellow: 0.2126(255) +0.7152(255) = 236.589
Luminance Magenta: 0.2126(255) +0.0722(255) = 72.624
Luminance Cyan: 0.7152(255) +0.0722(255) = 201.147

Figure 4: Luminance values from the three selected colour (Yellow, Magenta and Cyan)

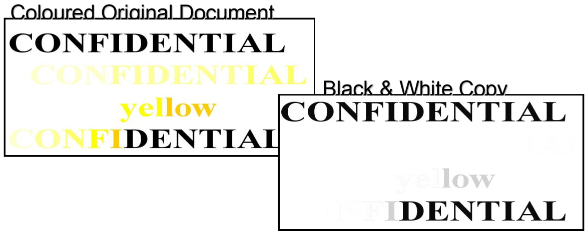


Figure 5: Left is the original coloured source and right is the grayscale copy

3.4 Strategic Watermark Location: Watermark Attack Prevention

Commonly, the word "Confidential" is popularly inserted as watermark into copied document and it is usually placed diagonally across a document as depicted. It is not only for aesthetic appearance but it also covers more area of the document to prevent or reduce chances of modification of the watermarked documents. The proof that diagonally placed watermark decreases the chances of attacks can be calculated with a simple probability theory. According to classical definition of probability from *Probability Tutorial* as in equation (4), assume that P(E) is the difficulty of watermark attacks based from how many wordings / objects are covered by the "Confidential" watermark. As for n(E), it is the total of words/objects being covered and lastly, n(S) is the total of the sample space of the watermark and the words/objects. From Table 4, it is clearly proven that the watermark placed diagonally has higher probability of difficulty in attacking the watermark as well as modification of the documents.

$$P\ (E) = \frac{\text{Number of elements in 'E'}}{\text{Number of elements in sample space 'S'}} \qquad (4)$$

Table 4: *"CONFIDENTIAL" mark placed at various locations for probability calculations.*

Situation	Calculation
Placed at the side of documents and covers none of the wording to the document	Let say, N(S)=100 n(E)=0 n(S)=100 P(E)=n(E)/n(S)=0
Placed at the middle horizontally and covers 10% of the area	Let say, N(S)=100 n(E)=10 n(S)=100 P(E)=n(E)/n(S)=0.1
Placed at the middle diagonally and covers 30% of the area	Let say, N(S)=100 n(E)=30 n(S)=100 P(E)=n(E)/n(S)=0.3

4.0 RESULTS

As earlier mentioned, this particular type of detection is suitable in the working environment where grayscale photocopier is commonly available. In Figure 6 [Left], the template watermark is created before the insertion of the watermark. Figure 6 [Right] shows the watermarked document after grayscale photocopying where it is seen that the seal and the confidential word are distorted and invisible and hence will prove that the document is illegally obtained. Even though, it may able to be colour photocopied, the details inserted into the database and imprinted on the top of the documents should be able to track and detect unauthorized copies of the original documents.

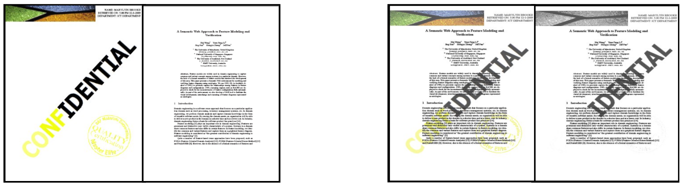


Figure 6: Left is the Watermark and right is the document source [Left Figure]. Left is the colour-printed hardcopy and right is the grayscale copy [Right Figure]

5.0 CONCLUSION

From the case study conducted, it is proven that yellow colour cannot be copied using grayscale devices. With the scientific calculation as the proof, it is also found that yellow has the second highest luminance after the colour white. Even though the printed documents can not be prevented from being copied, the implementation of the yellow watermark can trace or detect unauthorized or illegal copies of the printed documents from any grayscale photocopy environment. Even if the masqueraders try to assume ownership of the documents, the details imprinted on the document may prove otherwise. In addition, the manual stamping seal forms a part of the automation system for the Document Contorl System (DocCon) using PDF stamping technique discussed. As a result, the security for the documents is increased through DocCon as the digital watermarking technique is implemented (PDF stamping) and the softcopy of the files will never be retrieved by the users for any modification as the softcopy files are deleted as soon it is printed from the server side. In conclusion, the research shows that watermarking is one of the information hiding techniques which can be used to control and manage documents. Therefore, information hiding is rapidly becoming the most sought after method to protect industries in music, publications, filming, and other works which may be digitized. Even if the solution is provided for the grayscale photocopies, further enhancements will be done if better solutions are found from further researchs. The solution for the colour photocopies is yet to be accomplished further through research which will be more emphasized on the coloured copies of digital watermarkings. In addition, the embedding method for digital watermarking techniques should be thoroughly studied to enhance the security of the printed documents. As for the licensing problems encountered for the PDF stamping, various programming PDF stamping techniques are researched for better solution to the Document Control System (DocCon).

REFERENCES

Anonymous Author (2009). *Probability Tutorial*. Retrieved January 28, 2009, from http://www.tutors4you.com/probabilitytutorial.htm

Berns, R. S. (2000). *Billmeyer and Saltzman's Principles of Color Technology, 3rd edition*. Wiley, New York. ISBN 0-471-19459-X.

Bohren, C. F., and Clothiaux, E. E. (2006). *Fundamentals of Atmospheric Radiation: An Introduction with 400 Problems*. Wiley-VCH. ISBN 3527405038.

Chen, B., and Wornell, G. W. (2001). Quantization Index Modulation: A Class of Provably Good Methods for Digital Watermarking and Information Embedding. *IEEE Transaction on Information Theory*, May 2001, Vol. 47, No. 4, pp. 1423–1443.

Cox, I. J. Kilian, J., Leighton, T., and Shamoon, T. (1996). Secure Spread Spectrum Watermarking for Images, Audio, and Video. *Proceedings of the 1996 IEEE International Conference on Image Processing*, 1996, 3, pp. 243–256.

Dittmann, J., and Meg´ıas, D., Lang, A., and Herrera-Joancomart´ı, J. (2006). A Theoretical Framework for a Practical Evaluation and Comparison of Audio Watermarking Schemes in the Triangle of Robustness, Transparency and Capacity. *Otto-von-Guericke University of Magdeburg, Germany & Universitat Oberta de Catalunya, Spain*, 2006.

Fitzgerald, M.. (2004). *Using Javascript to Apply Template to a PDF File*. Byte Ryte, The Netherlands.

Gatter, M. (2005). *Getting It Right in Print: Digital Pre-press for Graphic Designers*. Laurence King Publishing.

Hirakawa, K., and Parks, T.W. (2005). *Chromatic Adaptation and White-Balance Problem*. IEEE ICIP.

Hirsch, R. (2004). *Exploring Colour Photography: A Complete Guide*. Laurence King Publishing. ISBN 1856694208.

Johnson, S. (2006). *Stephen Johnson on Digital Photography*. O'Reilly. ISBN 059652370X.

Kuehni, R. G. (2002). The Early Development of the Munsell System. *Color Research and Application*, *John Wiley & Sons, Inc.* , Vol. 27, Issue 1: pp. 20–27.

Kutter, M., and Petitcolas, F. A. P. (1999). A Fair Benchmark for Image Watermarking Systems. *Electronic Imaging '99. Security and Watermarking of Multimedia Contents*, January 25-27, 1999, 3657, pp. 1–14.

Martin Kutter (2003). *Digital WatermarkingWorld*. Retrieved January 28, 2009, from http://www.watermarkingworld.org

Meehan, J., Taft, E., Chernicoff, S., Rose, C. & Ron, K. (2005). *PDF Reference, Fifth Edition Version 1.6*. California: Peachpit Press.

Michael Stokes, Matthew Anderson, Srinivasan Chandrasekar, and Ricardo Motta (1996). *A Standard Default Color Space for the Internet - sRGB*. Retrieved January 28, 2009, from http://www.w3.org/Graphics/Color/sRGB

Ohno, Y., (1999). OSA Handbook of Optics, Volume III Visual Optics and Vision Chapter for Photometry and Radiometry. *Optical Technology Division*, Oct 20, 1999, pp. 1 – 17.

Poynton, C. A. (2003). *Digital Video and HDTV: Algorithms and Interfaces*. Morgan Kaufmann.

Ross, D. T., Goodenough, J. B., and Irvine, C. A. (1975). Software Engineering: Process, Principles, and Goals, *IEEE Computer*, Vol. 8, No. 5, May 1975, pp. 17 – 27.

Su, J. K. Hartung, F., and Girod, B. (1999). Digital Watermarking of Text, Image, and Video Documents. *Preprint submitted to Elsevier Preprint*, August 23, 1999, pp. 1–16.

van Schyndel, R. G., Tirkel, A. Z., and Osborne, C. F. (1994). A Digital Watermark. *Proceedings of the 1994 IEEE International Conference on Image Processing*, 1994, 2, pp. 86–89.

Welsh, W., Ashikhmin, M., and Mueller, K. (2002). Transferring Color to Greyscale Images. *Optical Technology Division*, 2002.

Wolfgang, R. B., and Podilchuk, C. I., and Delp, E. J. (1999). Perceptual Watermarks for Digital Images and Video. *AT&T foundation*, 1999 pp. 1-46.