Article

Metrics of automatic Image Quality Assessment based on Human Perception — A comparative study and a proposal of a new metric

André Neto [1,2,*], Advisors: Nuno Gonçalves [1,2], João Marcos [1,2]

[1] University of Coimbra, Department of Electrical and Computer Engineering, Coimbra, Portugal
[2] Institute of Systems and Robotics, Coimbra, Portugal
[*] Student Number: 2019237495

Abstract: The counterfeiting and trafficking of administrative documents have been longstanding issues, and more recently, we've seen an increase of fake visual content on the internet. As deepfake technology continues to advance at a rapid pace, there is a constant need to develop new methods to counteract its use. The VIS Team of the ISR-UC proposed various printer-proof steganography methods that aim to embed information covertly within an image, introducing somewhat perceptible changes that challenge the sensitivity of traditional image quality assessment metrics such as Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR), Learned Perceptual Image Patch Similarity (LPIPS) and Fréchet-Inception Distance (FID). As a consequence, these state-of-the-art methods may fail to accurately reflect the true quality of steganography-affected images. This raises concerns about the reliability of these methodologies in scenarios where content integrity is compromised. This article is a comparative study that focuses on studying how humans can very efficiently assess image quality and what is the underlying principal aspects of human judgement in steganography-affected content. Eventually proposing a human centered metric, so that the assessment of the image quality is in accordance with the generic human visual system and cognitive mechanisms.

Keywords: *Steganography; Image Quality Metrics; Face Image Quality Assessment; Human Perception of Image Quality*

1. Introduction

There has been a significant increase in the number of deepfakes detected globally across all industries from 2022 to 2023, with notable regional differences: 1740% deepfake surge in North America, 1530% in Asia-Pacific, 780% in Europe (inc. the UK), 450% in the Middle East and Africa and 410% in Latin America. Identity cards remain the most frequently exploited for identity fraud, accounting for nearly 75% of all fraudulent activities involving identity documents (IDs) [1]. Steganography adds one extra layer of security, acting like a signature on visual content.

Steganography [2,3] is the process of concealing a secret message covertly within an image, in such a way that this message is not visible to the human eyes. The main focus of steganography is to hide the fact that communication is occurring or that a secret message is being sent, ensuring a level of secrecy and confidentiality. As the amount of information (capacity) being coded increases, visual artifacts will become more noticeable, therefore we need image quality metrics (IQA) to control the trade-off between capacity and detectability. Moreover, when decoding physical printed images, there are two noise introducing phenomena to account for: the noise introduced by the machine that printed the image, and the noise that depends on the quality of the optical sensor (camera) that captures the image.

VIS Team of the ISR-UC has a research project with Imprensa Nacional-Casa da Moeda (INCM) for the encoding of hidden information in future generations of identity cards and

machine-readable travel documents (MRTDs). In order to decode encoded images present in this type of physical documents, it is required robust print resistant steganography techniques, due to the aforementioned two types of noise. The development of innovative, printer-proof steganography methods has posed a challenge for traditional IQA metrics. These metrics often fail to assess image quality as the general human would do.

In this article, we intend to understand how humans judge images and how replicable their judgement is, in order to develop a human centered image quality assessment metric that is closer to the common human perception. We'll start with a comparative study between traditional IQA metrics and collective subjective human perceptions of image quality using a set of identical images. The focus will firstly be in images of human faces but will eventually evolve to a more generic theme.

2. Related Work

In this section we'll introduce the metrics used in this study as well as the printer-proof steganography methods used to generate images with various visual qualities.

2.1. Image Quality Assessment Metrics

Image Quality Assessment finds application in fields spans from image compression, image enhancement to image restoration. IQA can be classified into objective and subjective types. Objective IQA is performed through computational models, and it can be divided into two categories: Full-Reference IQA (see Table 1), where the quality of a distorted image is assessed by comparing it to an undistorted version of itself; or No-Reference IQA, where the assessment of the image quality is done without any reference. In case of subjective IQA, human observers from the general public or domain experts assess the quality of the image and their collective rating yields a Mean Opinion Score (MOS).

Mean opinion score [4] has become a very popular indicator of perceived media quality. The International Telecommunication Union (ITU) has defined the opinion score as the "value on a predefined scale that a subject assigns to his opinion of the performance of a system" [5]. MOS is the average of these scores across subjects. Amongst its various areas of application, researchers have been using MOS to gather human judgments on the visual quality of images compressed using JPEG algorithms, contributing to its refinement and optimization over the years.

Table 1. State-of-the-art full-reference, objective IQA metrics.

Metric	Summary
Structural Similarity Index, SSIM	[6] measures the perceived structural information and patterns in images, taking into account luminance, contrast, and structure.
Peak Signal-to-Noise Ratio, PSNR	[7] is defined as the ratio of the peak signal strength to the root mean squared error (RMSE) between the original and distorted images.
Learned Perceptual Image Patch Similarity, LPIPS	[8] captures perceptual differences between images more in line with human visual perception. It utilizes a CNN based on VGG-19 [9] to assess the perceptual similarity between image patches.
Fréchet Inception Distance, FID	[10] is calculated by computing the Fréchet distance between two Gaussian fitted to feature representations of the Inception network [11].

2.2. Steganography Methods

A set of printer-proof steganography methods are briefly summarized in Table 2. These methods are typically based on Generative Adversarial Networks (GANs) [10] to encode and decode information. More recently, a noise simulation network was added to improve the ability to recover distorted images.

This is a novel challenge, the decoder should be robust to the unpredictable transformations that usually occur during the printing and digitization process, such as variations in contrast, the perspective of the acquired image, and color noise.

We can obtain various results depending on the method used. Figure 3 shows that there are noticeable differences in distortion amongst images produced using different steganography methods, but the same encoded information.

Table 2. State-of-the-art printer-proof steganography methods.

Methods	Summary
StegaStamp, 2020	[12] claims to be the first steganography model capable of decoding data from printed images. The authors show robust results in decoding data under physical transmission by developing novel strategies to add noise in the training process, printer noise simulation, and distortion for the training dataset.
CodeFace, 2021	[13] encoder and decoder networks are trained using end-to-end GANs. It introduces a new security system for encoding and decoding facial images that are printed in common IDs and MRTDs.
CodeFace with Riemannian Geometry, 2023	[14] proposes a new loss function that extends the loss function based on the L_2 distance between images to the Riemannian manifold of symmetric and positive definite matrices.
FStega, 2023	[15] proposes to employ Fourier Operators [16] to encode the message predominantly regions of medium to high frequencies of the cover image. Its encoder is based on the U-Net architecture [17], however, instead of using regular convolutions, they apply Fourier Neural Operator in each hidden layer. Also, to enhance the information recovery capabilities of the printed images, they propose an improved noise simulation process and create a decoder composed by several convolutional layers combined with a vision transformer.
StampOne, 2023	[18] focuses on high-level robust steganography, such as [12,13], striking a balance between high-quality encoded images and decoding accuracy. It mitigates distortion-related issues like JPEG compression, camera sensors and printer's Gaussian noise by incorporating gradient transform, wavelet transform, and Depthwise [19] to normalize and balance frequencies of the inputs.

3. Materials and Methods

The proposed method is based on the International Telecommunication Union Recommendation BT.500-15 [20] (ITU-R BT.500-15).

3.1. Methodology

The ITU-R BT.500-15 [20] provides methodologies for the assessment of picture quality including general methods of test, the grading scales and the viewing conditions. It recommends the double-stimulus impairment scale (DSIS) and the double-stimulus continuous quality-scale (DSCQS) method as well as alternative assessment methods such as single-stimulus (SS) methods and stimulus-comparison methods.

3.2. Dataset

In order to validate the proposed methods this work will depend on the construction of new datasets or the use of existing ones, like Labeled Faces in the Wild [21] dataset. The dataset images will be encoded with hidden messages using the steganography methods listed in Table 2. The images will be encoded with various levels of distortions. The focus will be firstly, in images of human faces, but will eventually evolve to a more generic theme.

A sample of the type of images in the dataset to be constructed can be seen in Figure 1.

Figure 1. Examples of the images in our dataset with different distortion levels.

3.3. Analysis of the results

A test will consist of a number of presentations, L. Each presentation will be one of a number of test conditions, J applied to one of a number of test sequences/test images, K. In some cases, each combination of test sequence/test image and test condition may be repeated a number of times, R.

3.3.1. Calculation of mean scores

The first step of the analysis of the results is the calculation of the mean score, $\bar{u}_{jkr}$ for each of the presentations:

$$\bar{u}_{jkr} = \frac{1}{N}\sum_{i=1}^{N} u_{ijkr} \tag{1}$$

where:

u_{ijkr}: score of observer i for test condition j, sequence/image k, repetition r

N: number of observers.

Similarly, overall mean scores, $\bar{u}_j$ and $\bar{u}_k$could be calculated for each test condition and each test sequence/image.

3.3.2. Calculation of confidence interval

When presenting the results of a test, all mean scores should have an associated confidence interval which is derived from the standard deviation and size of each sample.

It is proposed to use the 95% confidence interval which is given by:

$$\left[\bar{u}_{jkr} - \delta_{jkr}, \bar{u}_{jkr} + \delta_{jkr}\right] \tag{2}$$

where:

$$\delta_{jkr} = 1.96\frac{S_{jkr}}{\sqrt{N}} \tag{3}$$

The standard deviation for each presentation, S_{jkr}, is given by:

$$S_{jkr} = \sqrt{\sum_{i=1}^{N}\frac{(\bar{u}_{jkr} - u_{ijkr})^2}{(N-1)}} \tag{4}$$

With a probability of 95%, the absolute value of the difference between the experimental mean score and the 'true' mean score (for a very high number of observers) is smaller than the 95% confidence interval, on condition that the distribution of the individual scores meets certain requirements.

3.3.3. Screening of the Observers

First, it must be ascertained whether this distribution of scores for test presentation is normal or not normal using the β_2 test [22]. If β_2 is between 2 and 4, the distribution may be taken to be normal. For each presentation of the scores u_{ijkr} of each observer must be compared with the associated mean value, $\bar{u}_{jkr}$, plus the associated standard deviation, S_{jkr}, times two (if normal) or times $\sqrt{20}$ (if non-normal), P_{jkr}, and to the associated mean value minus the same standard deviation times two or times $\sqrt{20}$, Q_{jkr}. Every time an observer's score is found above P_{jkr} a counter associated with each observer, P_i, is incremented. Similarly, every time an observer's score is found below Q_{jkr} a counter associated with each observer, Q_i, is incremented. Finally, the following two ratios must be calculated: $P_i + Q_i$ divided by the total number of scores from each observer for the whole session, and $P_i - Q_i$ divided by $P_i + Q_i$ as an absolute value. If the first ratio is greater than 5% and the second is less than 30%, then observer i must be eliminated.

The above process can be expressed mathematically as:

- For each test presentation, we calculate the kurtosis coefficient, β_{2jkr}, where β_{2jkr} is given by:

$$\beta_{2jkr} = \frac{m_4}{(m_2)^2}, \quad m_x = \frac{\sum_{i=1}^{N}(u_{ijkr} - \bar{u}_{ijkr})^x}{N} \tag{5}$$

- For each observer, i, find P_i and Q_i.
  - If $2 \leq \beta_{2jkr} \leq 4$, then:

$$\begin{aligned} &\text{if } u_{ijkr} \geq \bar{u}_{jkr} + 2S_{jkr} \quad \text{then } P_i = P_i + 1 \\ &\text{if } u_{ijkr} \leq \bar{u}_{jkr} - 2S_{jkr} \quad \text{then } Q_i = Q_i + 1 \end{aligned}$$

  - else

$$\begin{aligned} &\text{if } u_{ijkr} \geq \bar{u}_{jkr} + \sqrt{20}S_{jkr} \quad \text{then } P_i = P_i + 1 \\ &\text{if } u_{ijkr} \leq \bar{u}_{jkr} - \sqrt{20}S_{jkr} \quad \text{then } Q_i = Q_i + 1 \end{aligned}$$

- If $\frac{P_i+Q_i}{J \cdot K \cdot R} > 0.05$ and $\left|\frac{P_i-Q_i}{P_i+Q_i}\right| < 0.3$ then reject observer i.

4. Work Plan

To achieve all the goals we proposed, it is important to have a work plan delimited on time and split into small assignments. For this dissertation, we will follow the Gantt Chart shown in Figure 2.

5. Preliminary Results

Following the steps on Section 3, we gathered five observers and applied the single-stimuli numerical categorical judgement method (SSNCS) from ITU-R BT.500-15, that uses an 11-grade numerical categorical scale, to rank the image quality in sessions of about 10 minutes each. The images in Figure 3 were evaluated by the observers.

5.1. Discussion

Table 2 shows the assessment given by the following metrics: SSIM, PSNR, LPIPS and the newly introduced MOS, applied to the images that result from the different steganography methods.

As we can see, objective IQA metrics tend to place images coded with CodeFace with Riemannian Geometry as the preferred method, followed by StampOne and then by FStega

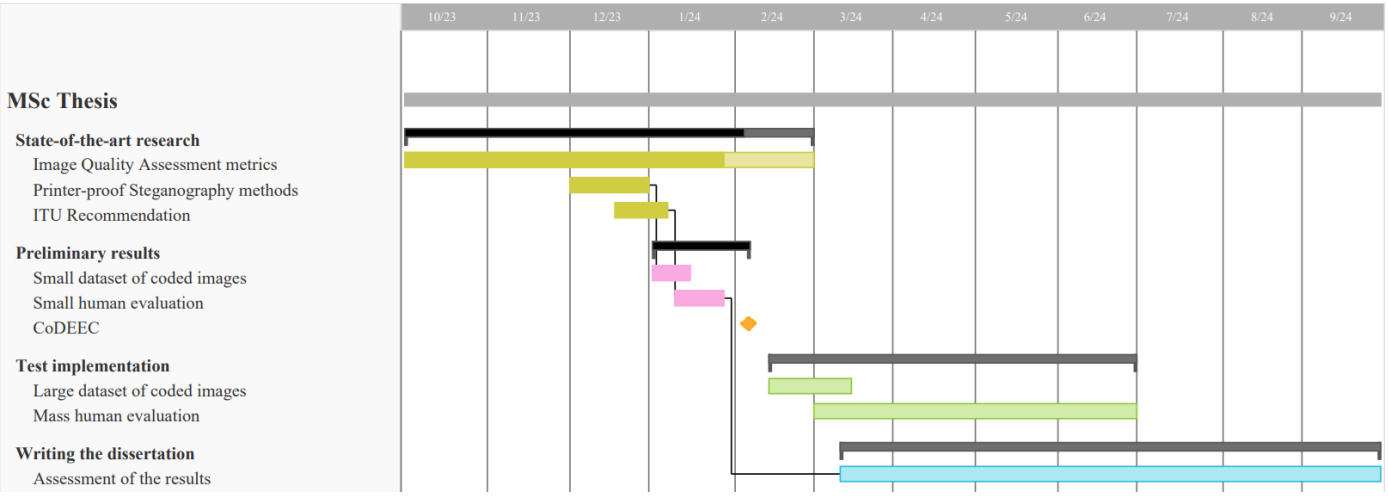

Figure 2. Proposed Gantt Chart for the Work Plan.

methods. However, observers show a clear preference for the StegaStamp method, which is placed substantially higher than any other method's subjective evaluation.

There are also some discrepancies between StampOne and FStega methods, as FStega gets better values for the LPIPS, and sometimes for the SSIM metrics, comparatively to StampOne, which is also a preferred method amongst observers.

5.2. Conclusion

There is a clear discrepancy between objective and subjective IQA metrics. This becomes very evident when the worst method selected by objective IQA for all subjects is the best method selected by observers.

5.3. Future Work

Within the scope of this thesis, in order to refine our results, we need to expand our dataset, and generate images with different levels of distortion. It is also necessary to create a controlled environment for the test section, and eventually test with more methodologies.

In order to propose a new metric, we need to know what factors create scenarios like ours, and study a way to approach an objective IQA to subjective evaluations.

5.4. Sampled Images

Figure 3. Which steganography method, from left to right respectively: Riemannian, StegaStamp, FStega and StampOne encoded images, best preserves the subject's identity recognition capability? In some cases, the human judgements disagree with state-of-the-art metrics.

Table 3. Metrics for each steganography method. For SSIM, PSNR and MOS, **higher** values are **better**. For LPIPS, **lower** values are **better**.

Subject A

Metrics	**Riemannian**	**StegaStamp**	**FStega**	**StampOne**
SSIM	**0.9670**	0.9335	0.9481	0.9574
PSNR	33.352	30.968	32.241	**34.871**
LPIPS	**0.0114**	0.0174	0.0172	0.0253
MOS	7.2	**8.2**	4.2	7.2

Subject B

Metrics	**Riemannian**	**StegaStamp**	**FStega**	**StampOne**
SSIM	**0.9667**	0.9274	0.9452	0.9302
PSNR	**33.240**	31.406	32.513	31.408
LPIPS	**0.0118**	0.0187	0.0203	0.0220
MOS	6.8	**8.2**	4.6	6.4

Subject C

Metrics	**Riemannian**	**StegaStamp**	**FStega**	**StampOne**
SSIM	**0.9762**	0.9448	0.9624	0.9575
PSNR	34.501	31.990	34.665	**34.706**
LPIPS	**0.0106**	0.0194	0.0160	0.0220
MOS	5.2	**6.4**	3.4	5.0

Abbreviations

The following abbreviations are used in this manuscript:

VIS Team	Visual Information Security Team
ISR-UC	Instituto de Sistemas de Robótica da Universidade de Coimbra
INCM	Imprensa Nacional-Casa da Moeda
IQA	Image Quality Assessment
HVS	Human Visual System
MOS	Mean Opinion Score
ITU	International Telecommunication Union
GANs	Generative Adversarial Networks
SSIM	Structural Similarity Index Metric
PSNR	Peak Signal-to-Noise Ratio
LPIPS	Learned Perceptual Image Patch Similarity
FID	Fréchet Inception Distance
SS	Single-stimulus
SSNCS	Single-stimulus Numerical Categorical Scale

References

1. Identity Fraud Report, 2023. A comprehensive, data-driven report on identity fraud dynamics and innovative prevention methods, © Sum and Substance Ltd UK.
2. Baluja, S. Hiding Images in Plain Sight: Deep Steganography. Advances in Neural Information Processing Systems; Guyon, I.; Luxburg, U.V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; Garnett, R., Eds. Curran Associates, Inc., 2017, Vol. 30.
3. Mandal, P.C.; Mukherjee, I.; Paul, G.; Chatterji, B. Digital image steganography: A literature survey. *Information Sciences* **2022**, *609*, 1451–1488. doi:https://doi.org/10.1016/j.ins.2022.07.120.
4. Streijl, R.C.; Winkler, S.; Hands, D.S. Mean opinion score (MOS) Revisited: Methods and Applications, Limitations and Alternatives. *Multimedia Systems* **2016**, *22*, 213–227.
5. IOO, G. Recommendation ITU-T P. 10 Vocabulary for Performance and Quality of Service. ITU-T, 07 **2006**.
6. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* **2004**, *13*, 600–612.
7. Yoo, J.C.; Ahn, C. Image matching using peak signal-to-noise ratio-based occlusion detection. *IET image processing* **2012**, *6*, 483–495.
8. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 586–595.
9. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition, 2015, [arXiv:cs.CV/1409.1556].
10. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, 2018, [arXiv:cs.LG/1706.08500].
11. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.
12. Tancik, M.; Mildenhall, B.; Ng, R. StegaStamp: Invisible Hyperlinks in Physical Photographs. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
13. Shadmand, F.; Medvedev, I.; Goncalves, N. CodeFace: A Deep Learning Printer-Proof Steganography for Face Portraits. *IEEE Access* **2021**, *9*, 167282–167291.
14. Cruz, A.; Schardong, G.; Schirmer, L.; Marcos, J.; Shadmand, F.; Gonçalves, N. Riemannian loss for image similarity-based comparison applied to printer-proof steganography. Submitted to 37th Conference on Neural Information Processing Systems.
15. Schirmer, L.; Schardong, G.; Cruz, A.; Shadmand, F.; Marcos, J.; Gonçalves, N. FStega: Fourier Neural Operators for Printer-proof Steganography. Submitted to the International Conference on Learning Representations, 2024.
16. Li, Z.; Kovachki, N.; Azizzadenesheli, K.; Liu, B.; Bhattacharya, K.; Stuart, A.; Anandkumar, A. Fourier Neural Operator for Parametric Partial Differential Equations, 2021, [arXiv:cs.LG/2010.08895].
17. Du, G.; Cao, X.; Liang, J.; Chen, X.; Zhan, Y. Medical image segmentation based on u-net: A review. *Journal of Imaging Science and Technology* **2020**.
18. Shadmand, F.; Medvedev, I.; Schirmer, L.; Marcos, J.; Gonçalves, N. StampOne: Addressing Frequency Balance issue in Printer-proof Steganography. ISR Technical Report. To be submited soon.
19. Tay, Y.; Dehghani, M.; Bahri, D.; Metzler, D. Efficient Transformers: A Survey, 2022, [arXiv:cs.LG/2009.06732].
20. IOO, G. Recommendation ITU-R BT.500-15 Methodology for the Subjective Assessment of the Quality of Television Pictures **2023**.
21. Huang, G.B.; Mattar, M.; Berg, T.; Learned-Miller, E. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. Workshop on faces in'Real-Life'Images: detection, alignment, and recognition, 2008.
22. Bai, J.; Ng, S. Tests for skewness, kurtosis, and normality for time series data. *Journal of Business & Economic Statistics* **2005**, *23*, 49–60.