Cover Page

1) Title of the paper:

To Recognize or not to Recognize - A Database of Encrypted Images with Subjective Recognition Ground Truth

2) authors’ affiliation and address:

LTeN, Polytech’Nantes, U6607 & RMeS UFR Odontologie, U1229
Tel : 02.40.68.31.56

3) e_mail address:

Florent.Autrusseau@polytech.univ-nantes.fr

4) Journal & Publisher information:

Elsevier Information Sciences
https://www.journals.elsevier.com/information-sciences

5) bibtex entry:

```
@article{InformationSciences2020,
     author = {H. Hofbauer and F. Autrusseau and A. Uhl},
     title = {To Recognize or not to Recognize - A Database of
     Encrypted Images with Subjective Recognition Ground Truth},
     journal = {Elsevier Information Sciences},
     year = {2020}
}
```

To Recognize or not to Recognize — A Database of Encrypted Images with Subjective Recognition Ground Truth

Heinz Hofbauer

Department of Computer Sciences, Paris Lodron University of Salzburg, Austria

Florent Autrusseau

Laboratoire de Thermique et Energie de Nantes & Regenerative Medicine and Skeleton, University of Nantes, France

Andreas Uhl

Department of Computer Sciences, Paris Lodron University of Salzburg, Austria

Abstract

The assessment of very low quality visual data is known to be difficult. In particular, the ability of humans to recognize encrypted visual data is currently impossible to determine computationally. The human vision research community has widely studied some particular topics, such as image quality assessment or the determination of a visibility threshold, while others are still barely researched, specifically visual content recognition. To this day, there does not exist a reliable recognition index that can be employed for such tasks.

In order to enable the study of human image content recognition, and in an attempt to propose a corresponding recognizability index, we build a dataset of selectively encrypted images together with subjective ground-truth about their human intelligibility. The methods of acquisition, setup, protocol, outlier detection, are described and we suggest how to calculate a recognition score as well as a recognition threshold. The performance of traditional visual quality indices to predict human visual content recognition is assessed on these data and found to be inapt to estimate recognition of visual content. Contrasting, structure based recognition indices as proposed for this task are shown to represent a promising starting point for further research. To facilitate the creation of a recognition index and to foster further research into human visual content recognition and its relation to the human visual system we will make the database publicly available.

Email addresses: `hofbauer@cs.sbg.ac.at` (Heinz Hofbauer), `Florent.Autrusseau@univ-nantes.fr` (Florent Autrusseau), `uhl@cs.sbg.ac.at` (Andreas Uhl)

Keywords: Selective Encryption, Human Visual Recognition of Low Quality Images, Visual Quality Indices, Database, Ground Truth, Recognition Threshold

1. Introduction

The general assessment of the security of encrypted visual data is difficult. In this context, security is understood as "visual security", i.e. the amount of visual information still present in such protected data, without considering cryptanalysis in the sense of analysing the cryptographic strength of the underlying cipher. In most partial / selective encryption schemes, the employed ciphers are beyond any doubt with respect to their cryptographic strength in any case, e.g. AES in some appropriate mode. What is of interest is the visual information still present in the data, in case of selective encryption in unprotected data parts and in the interference among protected and unprotected parts, respectively.

Selective encryption (SEnc) is the encryption of a *selected* part of a media file or stream, utilizing state of the art ciphers like AES. The goal is to achieve confidentiality for the visual content, or parts thereof, while still maintaining the file format. The latter means that the encrypted file is still usable as the media file it actually was before encryption. Specifically, format compliance means that a standard compliant decoder should be able to decode the encrypted image/video, and thus produce an output image/video. Thus, when we talk about the quality or recognizability of the encrypted visual data we talk about the output of a standard compliant decoder processing a format compliantly encrypted media stream.

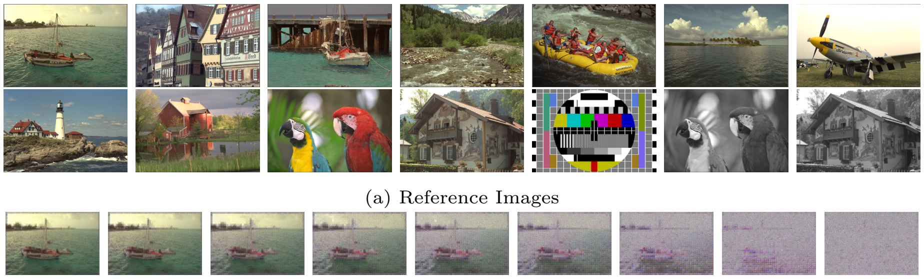
Over the years, a large number of SEnc schemes have been introduced, e.g., [4, 17, 36, 12, 5, 29, 28], and an essential question is how to assess their respective visual security. The unfortunate truth is that in most cases security is not properly assessed (see as example a discussion on methodology often employed to assess chaos-based encryption [26]). Often, authors only display a few images as 'proof'. Also, classical visual quality indices (VQI) are frequently used to estimate the visual security, e.g., PSNR and SSIM, while it has been shown that this is infeasible for the sufficient encryption scenario [10, 11], i.e., as soon as quality is very low and questions of intelligibility play a role. Even worse, it has been also shown that metrics originally intended to determine visual security also do not work properly [10].

When considering varying application context, visual security can have very different meanings [37]: *Transparent encryption* aims at revealing a low quality version of the content, e.g., as a preview to attract customers, while protecting the high quality version. *Sufficient encryption* aims to reduce the visual content to a level where a consumption of the image or video is no longer possible but does not care if potential content is leaked, e.g., consumers still recognize what is going on in a movie but the quality is so low that a pleasant viewing experience is prevented (pay-per-view scenarios). *Confidential encryption* is the next step where the goal is to actually make the content of the data unrecognizable. This is the type of visual security we target in this work. So the aim is to steer encryption strength not just to reduce quality, but to prevent scene and object recognition and understanding, respectively. Intelligibility of visual context has to be prevented. Note the difference to *cryptographical encryption* where the leakage of any data-related information has to be prevented which might enable to link plain and cipher text, respectively, not only considering visual clues.

For the first two of the above mentioned scenarii, the final user typically is a human observer in the context of a media entertainment context, controlled by digital rights management, staring at the protected images, and (eventually) being motivated by the reduced quality to pay for a full quality version. The question of identifying the visual content is either of no relevance (sufficient encryption) or is even a prerequisite of the application scenario (transparent encryption). For confidential encryption, the final user might be human as well as a recognition algorithm that tries to work on protected imagery. First work on the latter scenario to determine respective visual security has already been done: E.g. [18] considers the (dis)ability of SIFT-keypoint matching between images as a security metric and the (mis)conduct of proper segmentation is used as a security criterion in [19]. Also more application related investigations have been done by investigating the feasibility of biometric recognition on protected (i.e. selectively encrypted) data, e.g. in the context of fingerprint [2], iris [27], and fingervein recognition [32], respectively. Here, the security metric is the biometric comparison accuracy obtained.

In this work, we focus solely on human recognition in the context of confidential encryption. Application scenarios include to prevent leakage of movie content when distributing pre-release trailers or to prevent human security personnel from recognizing people in surveillance data. To automatically ascertain the recognizability of image content, the use of a visual recognition index (VRI), similar to a VQI, would be desirable. In order to come up with such a tool, similar to the process of establishing and assessing VQIs (compare the LIVE [31] and TID [24] datasets, respectively), a dataset with protected / encrypted visual material together with subjective ground-truth on the human intelligibility of these data is required. However, currently, neither proper VRI nor suited datasets to facilitate their developments do exist. In [34], a dataset of transparently encrypted images together with human perception ground-truth data has been released. However, this data is entirely unsuited for the human recognition context and VRI development (as transparently encrypted imagery is required to be recognizable by the application context). Also, the usage of the lower quality spectrum of impairments of the LIVE and TID datasets as done in [7] is of no use for the confidential encryption scenario as the intelligibility of the visual information in these two datasets is out of question.

Therefore, in this work, we intend to provide a corresponding dataset of encrypted visual data to facilitate the development of a VRI. The provided data also comprises results of human observer experiments at the recognition threshold of image content. Currently, as preparatory work, there only exists a set of recommendations for the acquisition of such a database, [6], which were adapted



(a) Reference Images

(b) Example of the *jpg* encryption type and all the steps from clearly recognizable to unrecognizable

Figure 1: Sample of Images contained in the database.

from the International Telecommunication Union (ITU) and International Electrotechnical Commission (IEC) recommendations for image quality evaluations to meet the confidential encryption context.

The rest of the paper is structured as follows: Section 2 describes the database, how the images were created and how the acquisition of subjective recognition scores were performed; Section 3 gives details on how to analyze the raw observer scores from the previous section, how outliers are found and how the content recognition threshold is estimated; Section 4 uses the database to test the capabilities of visual quality indices (VQI) to separate recognizable and unrecognizable images. Section 5 finally discusses the findings, gives guidelines and recommendations, and concludes the paper.

2. Database—Acquisition and Content

2.1. Images in the Database

One major aim when selecting target images for the inclusion in our database is that we wanted to have a considerable overlap with a database from literature, the LIVE database [31]. We have thus selected a set of images from the Kodak database[1] which are contained in the LIVE database as well. We also included an additional test image to have a non-natural image containing large single color patches, which can for example be found in drawings or animated films, and a more structured representation of frequency. To gauge the effect of color on the recognition we additionally included two gray-scale versions of images in the database. Overall, 12 color images, and 2 gray-scale ones were used. Despite the reduced number of images, the experiments had to be split into two sessions to combat viewer fatigue. Figure 1a shows the reference images contained in the database.

In order to maintain the experiment within a reasonable time-frame, we have selected 6 encryption methods operating with 9 protection strength steps

[1] `http://www.r0k.us/graphics/kodak`

(overall, 10 images including the original). The range was chosen "by hand" to include at least a clearly intelligible, albeit strongly distorted, version of the image and then in distinct steps towards an obviously unrecognizable image. This was done by selecting a parameter set for the encryption manually based on a single image and then applying this set of parameters to the whole database. The results for the *jpg* (described later) encryption can be seen in Fig. 1b.

The choice of encryption techniques was a difficult one and was limited by practical means. The encryption had to have settings allowing for a gradual change to get roughly 9 different levels of quality as required. Further, we used publicly available encryption methods only for reproducibility and ease of use. Even so, we had to "cheat" a bit with some encryption methods by iterating the application of the encryption to push it beyond the recognizability boundary.

Let us now provide some details on the six encryption types being used in this database:

jpg: This method [40] encrypts, based on AES, the AC coefficient coding of JPEG encoded images. The order of run-length coded symbols, and corresponding values, are permuted, as are coefficient value bits. In addition, the order of blocks using the same Huffman table with an interleaved minimum coded unit are permuted. The parameters for this encoding are either 'low' or the number of rounds. For low, only the first 8 AC coefficients were encrypted to get a proper high quality representation of the image. For the rest, multiple rounds of JPEG encoding and encryption were used since a single round of encryption, even if encrypting all AC coefficients, could not push the image beyond the recognition threshold.

H.265: The approach in [12] focuses on the encryption of sign coefficients in HEVC data. The encryption is done by creating a YUV sequence of length one (no motion) and applying the encryption to that sequence. The decoded frame is used as an encrypted image. In the context of H.265 this means we only encrypt I-frames. Since a single frame encryption has a rather limited impact on quality we used an iterative method similar to **jpg** to lower the quality, reencode and encrypt again for N rounds (an image is converted to YUV, encoded to H.265, encrypted and decoded to a single frame – this is one round). The parameters were split into two parts: what is encrypted and for which target quality. Quality is given either as the quantization parameter, or the number of rounds, in which case the quantization parameter was fixed at 22.

j2k* and *j2kne: The method from [35] encrypts a JPEG2000 file in either layer or resolution progression by encrypting CCPs (codeblock contribution to packet) of code blocks while maintaining signal markers. The ordering of contributing data in the bit stream is arranged such that the base information comes at the beginning, followed by refinement blocks which bring more and more detail into the image. Consequently, the encryption method can encrypt a window in this bit stream with an offset and a size. The difference between *j2k* and *j2kne* is that regular *j2k* uses error concealment in decoding, which tries to improve the image quality in case "strange" code block content arrives (which is the case for encrypted parts), while *j2kne* has turned error concealment off.

jxr: A set of format-compliant encryption methods for the JPEG XR standard is proposed in [17]: Coefficient scan order permutation, sign bit encryption, transform-based encryption, random level shift encryption, index-based VLC encryption, and encrypting entire frequency bands. A diverse set of parameters was used to get to desired steps in quality, specifically a combination of set sign bit encryption modes for DC sign bits, LP sign bits or HP/FLEXBITS. We also enable alternative transform encryption and random level shift encryption can be enabled for the DC, LP and HP bands. Here, we can specify the amount of encryption (as a percentage).

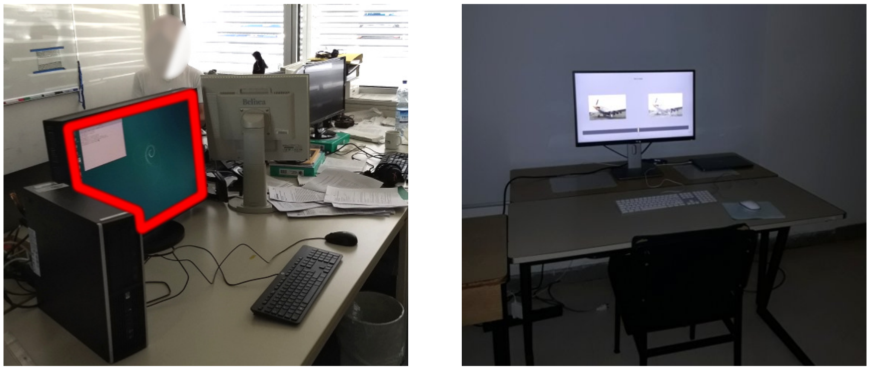
fake: This method is not actually an encryption method but a 'fake' encryption in the sense that it only spreads out information in the image thereby reducing its quality. The parameters give the operations used to transform the original into the "encrypted" version. The following operations are applied:

- generate an average map A with a varying window size X;
- scale down all pixel intensities I towards A with factor X, that is $I' = A + X(I - A)$;
- scale down all pixel intensities I towards A with factor uniformly distributed between 0 and X, that is $I' = A + \mathcal{U}(0, X)(I - A)$;
- noise is generated averaging pixel intensities I towards a uniformly distributed intensity;
- noise is generated averaging pixel intensities I towards a Gaussian distributed intensity.

Split of the Database for Acquisition: To prevent viewer fatigue we aimed to keep a single session of ground truth acquisition well below the one hour mark, in order to achieve this we split the database into two test sets: *SPLIT1* contains the encryption types *fake*, *j2k* and *j2kne*, *SPLIT2* is composed of *jxr*, *jpg* and *H.265*. This means a session should take at most 50 minutes : 14 images with 9 distortion steps for each of the three encryption types and maximum 8 seconds per image leads to 50 minutes and 24 seconds. In practice most sessions stayed well below this time since most pairs in the recognizable range can easily be selected faster than the allotted 8 seconds.

2.2. Setup for Recognition Groundtruth Acquisition

The Video Quality Experts Group (VQEG) and ITU groups regularly issue recommendations [39, 14, 1, 3] concerning the subjective setups and protocols for quality assessment. These recommendations are aimed towards the quality estimation based on human observer scores, or the visibility threshold tracking, but none are issued for a recognition task. In [6] these recommendations were evaluated in terms of applicability for the acquisition of a database for image content recognition. We will follow these combined guidelines [6] and adapt them where necessary. In the following we will briefly describe the environment and procedure used during the acquisition of the subjective recognition data.



(a) Setup for the tests in *SE*.

(b) Setup for the tests in *CE*.

Figure 2: Recording conditions for the two environments.

It should be noted that the subjective acquisitions were collected in two different locations: One had the option to have a more stringent setup (a controlled environment *CE*), the other only to a lesser degree (semi-controlled environment *SE*). The setups are shown in Fig. 2, in *CE* the room was properly darkened and had a controlled ambient light while in *SE* the blinds were drawn and an old monitor was used as a glare shield to prevent reflections on the test monitor.

Layout of the experiment: Three originals and three encrypted images are shown. The way this is set up and displayed to the user is shown in Fig. 3. One pair of images is an original and derived encrypted image, the other four images are unrelated. The participant must select the matching pair by clicking on the two related images. This is an adaptation of the two alternative forced choice protocol from the recommendations. The difference is that a larger possible number of choices delimit the non-recognition threshold better since the chance for a random 'correct' guess is lower. See the following sections for the calculation of the threshold and the impact of the probability on the calculation. We will refer to this method as **Match2** as this is what it was called in [6] where it was first recommended.

Viewing distance: In both the *CE* and *SE*, a supervisor instructed the observer to keep a proper viewing distance. The viewing distance was set to 6 times the images' height.

Scaling: Controlled and semi controlled environment have screens which are sized so that the images are not scaled.

Illumination and Calibration: Optimally a controlled environment and a high quality calibrated monitor is recommended. The specifications in the *CE*[2] were: Illuminant white point CIE D65, maximum screen luminance of 200 cd/m^2, screen gamma function of 2.20, contrast ratio/ black point of 2 cd/m^2. Moreover, the room background illumination was set to 10 lux. Our setup was

[2] Calibration was conducted with an *X-Rite i1 Display Pro* ©

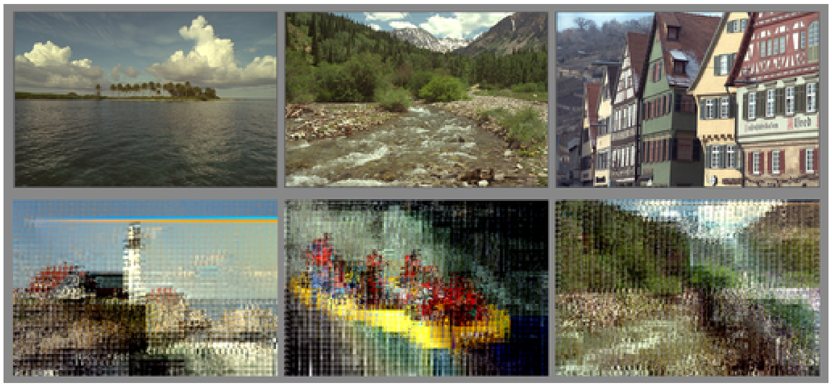


Figure 3: Setup of the acquisition experiment with three 'original' and three 'encrypted' choices.

thus compliant with the recommendations of [38], [15] and [16]. In *SE* these parameters were not controlled.

Viewing Time: We restricted viewing time to prevent observer fatigue and to ensure a timely conclusion, respectively, which is important for the acquisition of a large amount of data. The time chosen was 8 seconds in opposition to the recommended 10 sec [15]. The reasons for this are twofold: 1) the recognizability framework is easier than quality assessment and consequently takes less time and 2) it allows for more comparisons before observer fatigue sets in which is an important practical consideration.

Vision Check: For the *CE* environment a proper vision test was performed: Observers were screened to ensure perfect visual acuity and to detect possible color deficiencies. The Snellen eye chart was used to control the acuity, and Ishihara color plates were used to validate a normal color vision. For the *SE* setup the means were more limited but we utilized an online vision[3] test to check visual acuity, near vision and color vision.

About the Observers: The minimum number of observers recommended by all standards is 15 and was exceeded in all environments. In *CE*, 90 observers took part in the study with 45 observers enrolled per half session (*SPLIT1* and *SPLIT2*, see Split of the Database for Acquisition in Section 2). As previously mentioned, the observers were screened for correct color vision and acuity. Except for 3 observers who had a 20/25 vision, all other observers had at least a 20/20. Three observers had a red/green color deficiency. The observers' average age was 33.6 and 34.8 for *SPLIT1* and *SPLIT2* respectively. Only one of these observers (with a 20/25 acuity) was discarded during the outliers detection step (as explained later in section 3).

In *SE*, 60 Observers took part in the study with 30 per half session. The online test revealed two observers with a red/green deficiency. The average age

[3] https://www.essilor.com/en/vision-tests/test-your-vision

was 35.37 and 34.62 for *SPLIT1* and *SPLIT2* respectively.

During the subjective experiment, observers were asked to use their contact lenses or spectacles. Two distinct populations were tested. Most of the observers enrolled for *CE* were not computer scientists, but came from biology departments within the University of Nantes, whereas most of the observers of *SE* were computer scientists from the University of Salzburg, and some were familiar with the input images and distortions.

2.3. Where to get the Database?

The database is available online as University Salzburg Encryption Evaluation Database (USEE DB) at `http://www.wavelab.at/sources/USEE`. The database contains the original images, encrypted images and the individual binary, correct/incorrect match, and score per user (the output of the experiments). It also contains a score for each image based on the analysis methods in the following chapters and likewise a classification into recognizable or hidden image content. The encryption parameters for each file is explained in the database documentation.

3. On the Calculation of the Recognition Threshold

3.1. Techniques for the Analysis of Data

The data generated in the experiments differ from common quality experiments outputs. The main difference is that during quality evaluation each observer gives a numerical (quality) rating for each image, and based on these ratings the outliers can be detected. For each recognition task conducted here, the observer only generates a binary output, either “content recognized” or “content not recognized”. The final score for each recognition task is an aggregate over all observers, and is expected to trend towards the probability of random choice in the case of unrecognizability.

Outlier detection: We will follow the guidelines and reasoning laid out in [6] regarding outlier detection, and we will briefly summarize the relevant parts here. Outlier detection in the classical sense will not work, since the data being collected are not numerical scores, but rather binary decisions representing correct or incorrect recognition. A simple error aggregate also will not work since two observers can have the same number of errors while not agreeing on a single image. The solution is to view results as a vector with binary values, then the Hamming distance (HD), that is the number of differences between the two vectors, can be used to compare two observers. An outlier is an observer with a large distance from the majority of observers. We perform a hierarchical clustering which starts with the smallest distance and continues to cluster the elements (i.e. observers) together until a single cluster has formed. The outliers can then be detected based on statistics of similarity between observers: with O being the set of observers and

$$D = \{\mathrm{HD}(O_i, O_j) \,|\, \forall O_i, O_j \in I, i \neq j\} \tag{1}$$

the set of pairwise distances, we use the z-score

$$z_D = \mu(D) + 3\sigma(D), \tag{2}$$

where μ is the expected value and σ is the standard deviation, to find observers which are very far from the group consensus. As aggregation of cluster size, the maximum over all pairwise distances is used, meaning that all pairwise distances in the cluster are below the chosen threshold $z_D = \mu + 3\sigma$.

The way the result of this analysis looks in practice can be seen in Figure 4 later in the paper. The clustering is displayed as a dendrogram, a tree of merge decisions, the y-axis gives the merges at the height of the new cluster size. If the tree is cut off at a height matching z_D it will split into sub clusters where each cluster does not contain outliers. The largest such cluster is then used as the “correct” set of observers, the others are considered outliers (marked in red in the dendrograms).

Calculation of a score and ordering by recognizability: We have a set of numbered images $I_i \in I$, $i = 1, \ldots, \#I$ to be evaluated by a numbered set of observers, $O_j \in O$ for $j = 1, \ldots, \#O$. Each observer is represented by a vector containing one decision per image, either recognized (a match is found) with a score of 0 or not recognized with a score of 1: $O_j = (o_j^1, \ldots, o_j^{\#I})$, $j = 1, \ldots, \#O$ with $o_j^i \in \{0, 1\}$. A recognizability score for an image $S(I_i)$ can then be calculated as:

$$S(I_i) = \frac{1}{\#O} \sum_{j=1}^{\#O} o_j^i. \tag{3}$$

An ordering based on recognizability can now be established based on $S(I_i)$, we also know that for unrecognizable images $\lim_{\#O \to \infty} S(I) = 1 - p_*^r$, $*$ depending on evaluation method and P_*^r being the probability of randomly getting a recognition result if the images are unrecognizable. In our case the probability of a random correct guess is one in three for the original and one in three for the encrypted version, thus $p_{\mathbf{Match2}}^r = \frac{1}{3}\frac{1}{3} = 0.1\dot{1}$, compare Fig. 3. With the typical number of observers being at the maximum in the hundreds the limit will not be reached and we have to use a stochastic threshold.

We want to set a threshold T such that an image I with $S(I) > T$ is considered unrecognizable and less than 5% of unrecognizable images should be counted as recognizable.

The probability for an unrecognizable image U to have V recognizable decisions, and a consequent score of $S(U) = \frac{\#O - V}{\#O}$ is the number of possible vectors of that form (c_V) times the probability of such a vector (p_V), with

$$c_V = \binom{\#O}{V}, \tag{4}$$

$$p_V = {p_*^r}^V (1 - p_*^r)^{\#O - V}. \tag{5}$$

The probability for no more than V recognizable decisions, that is $S(U) \geq$

Table 1: Examples of the threshold (T), the number of incorrect recognition decisions (V_T) to reach T and the probability to reject an unrecognizable image as recognizable ($P(V_T)$) at that threshold are given for the experimental setup.

$\#O$	$p^r_{\mathbf{Match2}} = 0.1\dot{1}$		
	V_T	$P(V_T)$	T
50	9	95.40 %	0.82
100	16	95.07 %	0.84
500	67	95.24 %	0.87
1000	128	95.77 %	0.87

$\frac{\#O-V}{\#O}$ can be calculated as

$$P(V) = \sum_{i=0}^{V} c_i p_i = \sum_{i=0}^{V} \binom{\#O}{i} {p^r_*}^i (1 - p^r_*)^{\#O-i}. \tag{6}$$

Now we can calculate the threshold T, and corresponding V_T such that $P(V_T) > 0.95$ (that is the chance that a non recognizable image is counted as recognized by chance is less than 5%). Resulting example thresholds, and according values for V_T and $P(V_T)$ for varying $\#O$ are shown in Table 1.

Note that this boundary is based on the probability for a random guess in case the image is unrecognizable. This is because we do not have the probability of a correct assessment for any other case, i.e., when some information is retained in the image. This means that when we set the threshold to only miss 5% of the images which are unrecognizable, without taking recognizable images into account (due to unknown probability). Since there is an overlap this means that the higher the $P(V_T)$ the higher the (unknown) chance, and therefore number, of images being categorized as unrecognizable which are recognizable.

3.2. Analysis of Forced Choice Subjective Experiment Data

3.2.1. Outlier Detection

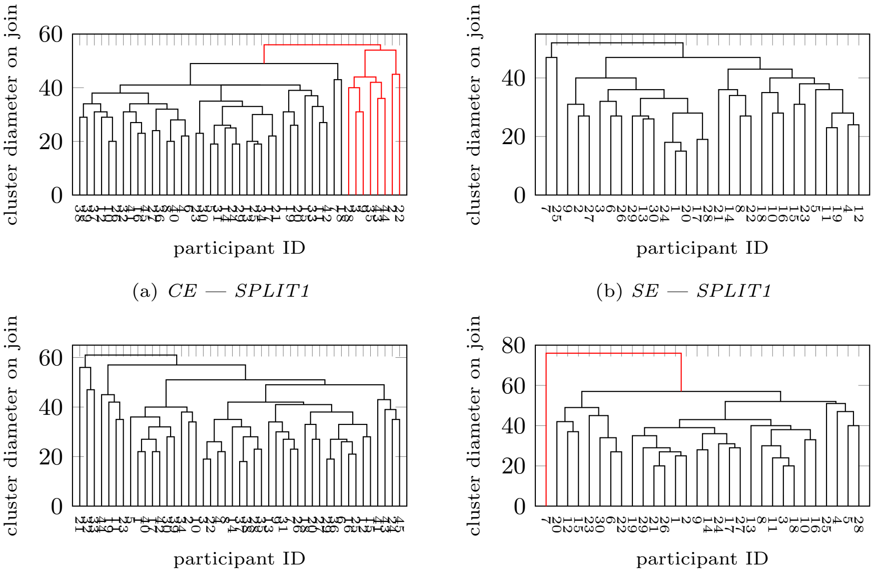
Outlier detection is performed per session and per environment based on the clustering technique and opinion difference expressed by Hamming distance as described earlier. This is done since, due to data retention policy, anonymised IDs from one set of experiments could not be linked to those from another. The results of the outlier detection are given in Table 2, we give the number of outliers as well as the intermediate values, μ and σ, (which give the distance between pairwise observer results) and the resulting cut off threshold T.

The experiments were not of the same difficulty, as exhibited by the mean number of errors over all observers – for *SPLIT1*, this is 76.22 and for *SPLIT2* it is 48.55. The dendrogram representations of the clustering are given in Figure 4.

Interestingly, we find that the difference between the two environments overall is small. Table 2 shows similar values for average and spread of scores, this can also be seen from Fig. 5a and 5b which show a direct comparison of scores.

Table 2: Distribution of observer difference and resulting threshold T and number of outliers for the experiments.

Setup	Testset	μ	σ	T	Outliers
CE	*SPLIT1*	34.74	6.72	54.89	8
SE	*SPLIT1*	34.05	6.18	52.58	0
CE	*SPLIT2*	37.36	8.50	62.85	0
SE	*SPLIT2*	38.80	8.83	65.30	1



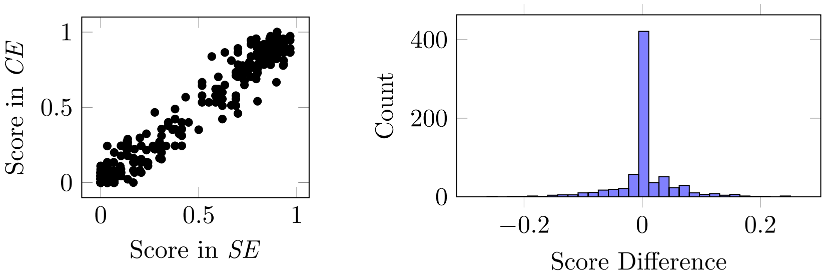
(a) *CE* — *SPLIT1*

(b) *SE* — *SPLIT1*

(c) *CE* — *SPLIT2*

(d) *SE* — *SPLIT2*

Figure 4: Dendrograms of the hierarchical clustering, outlier branches are shown in red.



(a) Scatterplot

(b) Histogram

Figure 5: Differences in the recognizability score from *CE* and SBG.a

When the threshold calculated from the more constrained environment is used for outlier detection in the SBG datasets the same outliers are detected. So a slightly less constrained environment seems to be perfectly fine for acquisition of data (which significantly reduces the efforts to conduct the subjective recognition experiments).

The final recognition score S is calculated based on the averaged outlier prune individual scores. However, we can also look at the scores per setup to see the difference in the results between mostly computer scientists (*SE*) and mostly non-computer scientists (*CE*). The recognizability score per setup is plotted as a scatter plot in Fig. 5a, the x-axis is the score from *SE* and the y-axis from *CE*. If the two scores agree the data point will be on the prime diagonal, the farther displaced from the diagonal the higher the disagreement is. The difference in scores is capped, the points are all in a band around the diagonal, which is a result from the outlier removal of course. What can not be seen is the distribution of differences: A point is plotted in the scatter plot if that particular combination occurs, no matter how often this is the case. Figure 5b therefore gives a histogram of difference scores, in 31 bins. And here we finally see that the disagreement becomes less frequent the stronger it is: In terms of Fig. 5a this implies that data points with a higher distance from the diagonal are less frequent observed.

3.2.2. Recognition Threshold

The calculation of the threshold is dependent on the number of observer scores taken into account. From Table 2 we can see that *SPLIT1* has a different number of outliers than *SPLIT2* leading to a different threshold for each set of tests. The matching protocol using 3 original and 3 distorted images has the probability of randomly selecting the correct pair of $p^r_{\mathbf{Match2}} = 0.1\dot{1}$, leading to the thresholds: $T = 0.821$ for $\#O = 67$ (*SPLIT1* containing *fake*, *j2k* and *j2kne*) and $T = 0.824$ for $\#O = 74$ (*SPLIT2* containing *jxr*, *jpg* and *H.265*). This was calculated following the method outlined earlier.

Figure 6 gives examples for each encryption type just below and just above the threshold, the same base image was used in each row. It should be noted that the recognizability scores are not really distributed equally, rather most images

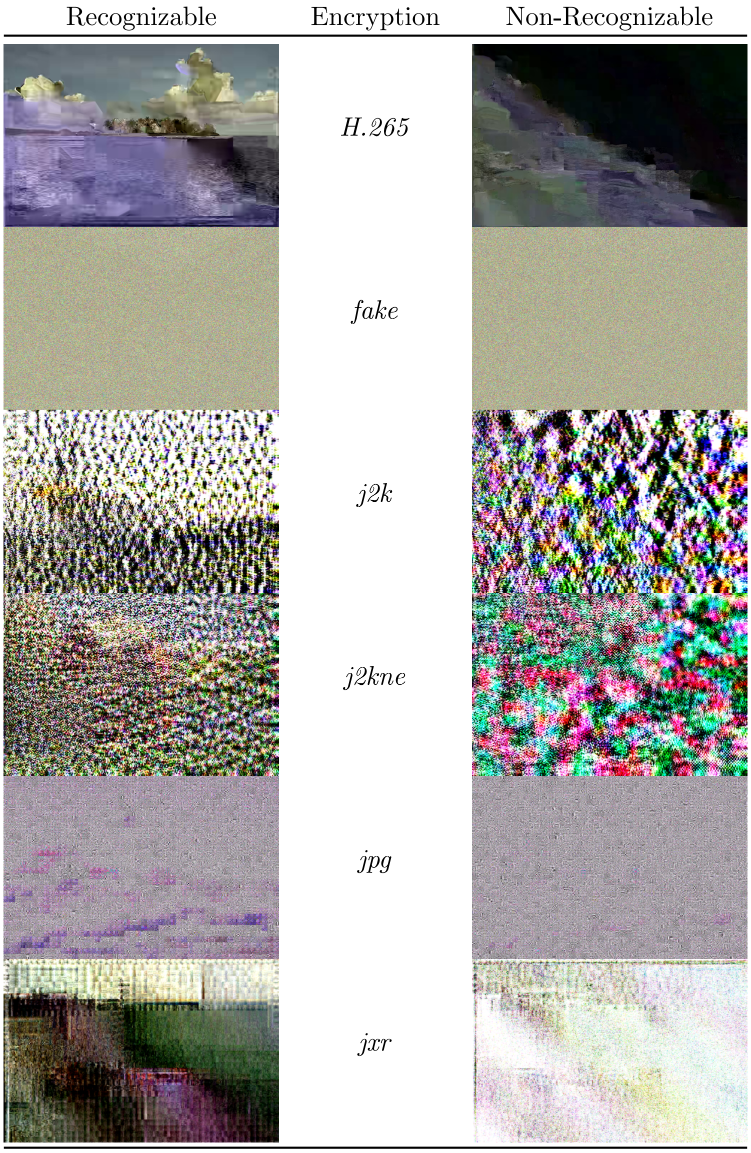


Figure 6: Example images which are just below and just above the recognition threshold.

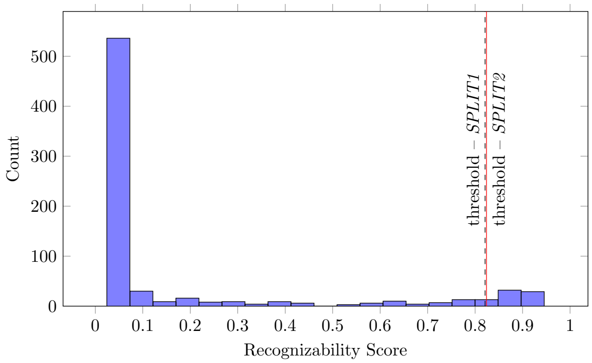
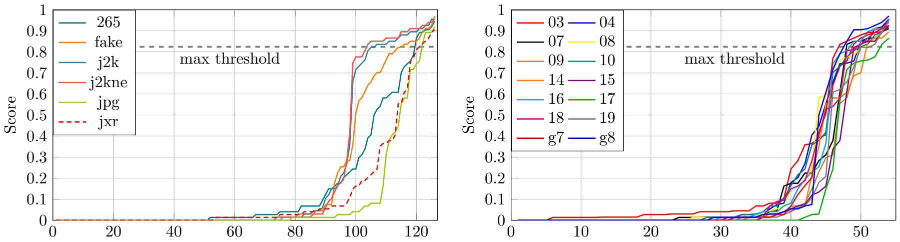


Figure 7: Histogram of recognizability scores, separated into 20 bins. The recognizability thresholds for *SPLIT1* and *SPLIT2*, which separate recognizable (below threshold) and unrecognizable, are also given.



(a) Grouping by encryption type.

(b) Grouping by image.

Figure 8: Images are split by encryption type or reference image, sorted based on the recognition score and plotted.

are clearly recognizable or clearly non-recognizable. A few images have scores in the middle range, where observers are split between recognition and non-recognition. However, there are only few images with these middle-range scores and most images can clearly be classified. A histogram of the recognizability scores is plotted in Fig. 7. The thresholds for both sets are given as well, the difference is rather small even given the difference in outliers.

3.3. Range of Recognition Scores

The database contains images with a recognition score and a classification, based on the threshold, into recognizable and non-recognizable.

The intent when setting the distortion steps was to produce a set of encrypted images which span the range from clearly recognizable to clearly unrecognizable for each reference image.

If we have met our goal according to the judgment by human observers can be seen in Fig. 8. The figure contains two subplots with different groupings, one by encryption type and one by source image. The images are sorted based on the recognition score with every point on the x-axis being one image. The threshold is given as the maximum of *SPLIT1* and *SPLIT2* thresholds. The goal was to create a spread of images (for each encryption type and source image) which span the range from recognizable to non-recognizable. Looking at the figure, we have achieved this goal. Both sub graphs show that the respective grouping has images over the threshold, that is for each distortion type we span the range and likewise for each source image we have results from recognizable to non-recognizable.

These graphs also confirm findings from before and illustrate them nicely. The first is that the score mostly falls into the broad categories of 'clearly recognizable', i.e. a score below 0.1, or 'not recognizable', i.e. at or above the threshold. This is reflected by the steep incline in the figures, i.e. very few images actually have a score between these two extremes. The other confirmation is that *SPLIT1*, specifically *j2k*, *j2kne* and *fake*, are more difficult to assess, presumably due to the encryption artifacts being more disruptive to the human visual system. We assumed this to be based on the larger σ during outlier calculation, reflecting observer disagreement. In Fig. 8a it can be seen that these encryption types have a larger number of images over the threshold. Here we have more results of 'random' selection during the **Match2** protocol, leading to a higher disagreement between observers, i.e. they did not guess alike, which was the design goal of the **Match2** protocol.

4. Experiments on Exemplary Usage of the Database

4.1. Methods for Assessment of Visual Recognition Index Candidates

The utilization of human observers for evaluation of recognizability is typically inpractical in real applications, so the employment of a visual recognition index (VRI) is highly desirable. To evaluate potential VRIs, a common set of methods is equaly desirable.

There are two tools which are conceivable for the purpose of separating recognizable and non-recognizable images. One is a simple classification method, the other is a score based system. The score based system can either produce a recognizability score and a non-recognizability threshold or it gives a classification and a certainty score. For evaluation purposes these two are the same and can be dealt with in the same way.

The use of evaluation methods for classification techniques lends itself to the task, which are based on true and false positives and negatives, respectively. We denote recognizable images as positive since they need to be corrected to be unrecognizable in an encryption system.

While the system can always be tuned to not produce false negatives, by rating everything as positives, this has to be counteracted in a measure by also taking into account the false positive rate. An additional problem is the unbalanced number of recognizable and unrecognizable images. A good match, c.f. [25], for the described circumstances is Matthews correlation coefficient (MCC) [23] which is related to the chi-squared statistic for the contingency table.

$$\mathrm{MCC} := \sqrt{\frac{\chi^2}{n}} = \frac{tp \times tn - fp \times fn}{\sqrt{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}}, \tag{7}$$

with tp, fp, tn, fn being true positives, false positive, true negative and false negative respectively.

For score and threshold based methods the same, as for the classification methods, holds true when it comes to the classification end result. However, since the threshold can be set, the resulting classification can be steered. To properly reflect this the minimum MCC over all thresholds should be used when evaluating a score based methods, to facilitate comparison with purely classification based methods.

Otherwise, common operating points on the receiver operating characteristic (ROC) curve based on the full database should be used. Two points on the ROC curve make sense: 1) the equal error rate, because it makes different methods comparable by computing bounds for the significant difference without needing the exact experimental results [9]; and 2) the false positive rate at zero false negative rate (0FNR), which is basically the desired output of the system, i.e. no false negatives means no breach in security.

In addition to the classification accuracy provided by the MCC a more direct comparison of the scores from human observers and algorithms can be taken into account. Given that the Human Visual System (HVS) is non-linear and the evaluation tool is not required to be, a non-linear correspondence measure should be used, usually the Spearman rank order correlation [33] or the Kendall τ [20]. Both are well known and widely used, to stay in tune with [10] we will use the Spearman rank order correlation (SROC).

$$\mathrm{SROC} := \frac{\mathrm{cov}\left(rg(x), rg(y)\right)}{\sigma_{rg(x)}\sigma_{rg(y)}}, \tag{8}$$

where $rg(x) : \mathbb{R} \mapsto \mathbb{N}$ is a function returning the rank of the argument, cov is the covariance and σ is the standard deviation. There is a slight problem with the massing of '0' scores, i.e., very clearly recognizable images (see Fig. 7), which might mess with the correlation. To prevent any problems it is suggested to remove any images which are obviously recognizable, scores close to 0, from the comparison. Considering potential operator errors we assume an error rate of no more than 10% which would exclude images with recognition scores $S(I) < 0.1$. Reporting this $SROC_{90}$, spanning 90% of the recognition scores, is preferable over the full SROC. In addition to the previously discussed scores for the recognition task we will also give the Pearson correlation (r), a type of linear correlation which is like the SROC without the rank function ($r := \frac{\mathrm{cov}(x,y)}{\sigma_x\sigma_y}$), and the root-mean-square error (RMSE) as suggested by [41].

4.2. Visual Quality Indices and Recognition

There are currently no visual recognition indices as far as the authors are aware. However, visual quality indices (VQIs) have been frequently used for visual security analysis, although it is known that they are not well suited for the task [10]. According to [11] the best suited VQIs for security analysis are the VIF and LEG, the most widely used are the PSNR and the SSIM. In the following we will briefly describe the VQIs and evaluate their usefulness as VRI based on the method outlined in the previous section using the database introduced.

Local Edge Gradients (LEG) The LEG [8] analyses changes in luminance and edge information. The main part is the edge score which combines the edge change between images, based on the local binary pattern concept, and an edge gradient change score between images. The edge change is calculated in the low frequency band and the gradient change in the high frequency band of a wavelet decomposition of the image. The LEG is a quality metric, a high metric score reflects a high quality, with a normalized score in $[0, 1]$.

Peak Signal to Noise Ratio (PSNR) The peak signal-to-noise ratio (PSNR) is still widely used because it is unrivaled in speed and ease of use. The PSNR is a quality metric, meaning a high metric score reflects a high quality, which gives a score in the range $[0, \infty]$. However, it is also well known that the correlation to human judgment is somewhat lacking even for high and medium quality [13].

Structural Similarity Index Measure (SSIM) The structural similarity index measure (SSIM) [42] extracts three separate scores from the image and combines them into the final score. First the visual influence is calculated locally then luminance, contrast and structural scores are calculated globally. These separate scores are then combined with equal weight to form the SSIM score. The SSIM is a quality metric, a high metric score reflects a high quality, which gives a score in the range $[0, 1]$.

Visual Information Fidelity (VIF) For the VIF, [30], a refined model is used which starts with the modeling of the reference image using natural scene statistics (NSS). Furthermore, the possible distortion is modeled as signal gain and additive noise in the wavelet domain. Parts of the HVS which have not been covered by the NSS are modeled, i.e. internal neural noise is modeled by using an additive white Gaussian noise model. Using this model the VIF score reflects the fraction of the reference image information which can be extracted from the impaired image.

4.2.1. Evaluation of VQIs

Table 3 gives the resulting scores of the assessment techniques as discussed earlier. Since the VQIs produce a numerical score a threshold is required for the classification, thus we only reported the MCC at the threshold with the minimum MCC.

None of the tested VQIs offers acceptable performance for a recognizability application. This was indeed expected, as the VQIs are commonly designed to detect small signal differences located onto (or nearby) edges. The encryption, c.f. Fig. 6, destroys content efficiently, as it significantly damages the low frequency content, and introduces a large amount of high frequency noise.

Table 3: Evaluation of different VQIs on the database, reporting the equal error rate, false positive rate for zero false negatives, the maximum absolute Matthews correlation coefficient, the absolute Spearman rank order correlation, the root mean squared error, and the Pearson correlation for the full database as well as per encryption type.

(a) LEG

Testset	EER [%]	0FNR[%]	$\|MCC\|$	$\|SROC_{90}\|$	RMSE	r
all	28.62	100.00	0.373	0.267	0.730	0.359
jpg	17.64	100.00	0.427	0.180	0.766	0.383
jxr	16.46	100.00	0.537	0.388	0.733	0.487
265	28.36	85.71	0.369	0.192	0.688	0.206
j2k	16.77	70.00	0.667	0.505	0.718	0.474
j2kne	17.19	73.91	0.715	0.404	0.720	0.504
fake	12.91	81.82	0.592	0.513	0.750	0.483

(b) PSNR

Testset	EER [%]	0FNR[%]	$\|MCC\|$	$\|SROC_{90}\|$	RMSE	r
all	20.83	100.00	0.508	0.385	10.936	0.464
jpg	9.96	100.00	0.515	0.296	13.228	0.534
jxr	17.92	100.00	0.531	0.334	11.176	0.561
265	42.65	100.00	0.246	0.074	10.569	0.133
j2k	11.44	90.00	0.710	0.413	10.501	0.543
j2kne	12.10	82.61	0.774	0.416	10.511	0.547
fake	26.90	100.00	0.319	0.115	9.232	0.423

(c) SSIM

Testset	EER [%]	0FNR[%]	$\|MCC\|$	$\|SROC_{90}\|$	RMSE	r
all	29.17	98.61	0.375	0.321	0.722	0.288
jpg	6.24	100.00	0.781	0.520	0.820	0.336
jxr	7.92	83.33	0.608	0.557	0.662	0.503
265	44.33	100.00	0.246	0.099	0.512	0.266
j2k	34.72	90.00	0.320	0.085	0.804	0.337
j2kne	23.61	100.00	0.474	0.096	0.806	0.337
fake	11.60	72.73	0.654	0.604	0.677	0.538

(d) VIF

Testset	EER [%]	0FNR[%]	$\|MCC\|$	$\|SROC_{90}\|$	RMSE	r
all	18.51	95.83	0.460	0.304	0.832	0.326
jpg	19.71	100.00	0.290	0.102	0.900	0.359
jxr	16.46	50.00	0.698	0.462	0.857	0.323
265	14.08	57.14	0.644	0.660	0.829	0.448
j2k	19.67	95.00	0.545	0.277	0.817	0.410
j2kne	20.82	100.00	0.620	0.346	0.818	0.405
fake	20.49	72.73	0.570	0.387	0.767	0.445

VQIs are either purely statistical (such as the PSNR or MSE), and can easily be fooled by a high extent of noise in the images such as the one we encounter in this work, or are more advanced (VIF and LEG) and model the HVS behavior, and look for perceivable contrast modifications (i.e. nearby the images edges). Typically, advanced VQIs employ the Contrast Sensitivity Function (CSF) in order to represent the images as contrast gaps. The contour areas are thus presenting higher weights to be pooled into a single quality score. For the images composing this database, we have either high frequency (and high amplitude) noise on top of the image, or some large image areas being wiped out. An advanced HVS based VQI can simply not interpret the huge quantity of high contrast areas that occur in the distorted image and do not match with any existing contrast within the original image.

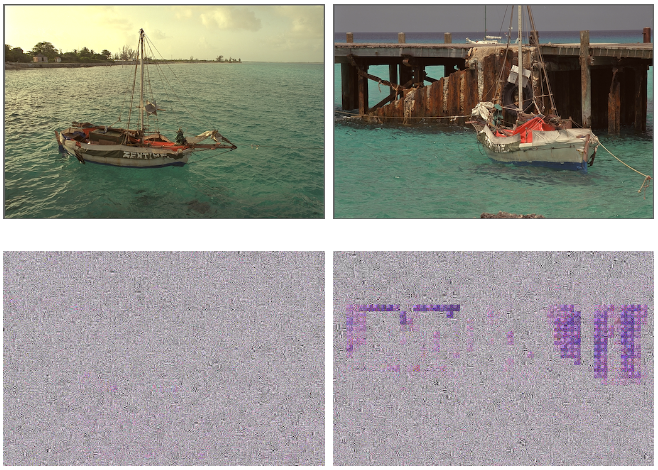
The recognition task is very different from a quality assessment or even from a visibility threshold detection task. For image recognition, an observer only needs to find a match between small portions of the images to recognize the content. The problem at hand is more similar to template matching than to quality assessment. Therefore, we will look at two more propositions: I) The task at hand is more like template matching and/or based on coarse structures of the image and II) the features of the VQIs are good but their combination (that is the way they are weighted) is bad. First (I) we will look at a template matching and structure based segmentation approach. Second (II) we will try to use the VQI scores on different resolutions and learn a scaleable vector regression model to see if the results can be improved.

4.2.2. Structure as Recognition Indicator

We use the 2D Normalized Cross-Correlation [21] as an estimate of the recognizability between the original and distorted images.

Normalized Cross-Correlation (NCC) Putting the human vision system perspective aside, we can see an image pair as two signals which may have some similarities. A common tool for evaluating the degree of similarity between two signals is the Cross-Correlation, which is basically a sliding dot product. The Normalized Cross-Correlation (NCC), being less sensitive to linear changes in the images amplitude has proved its efficiency in pattern recognition applications.

To illustrate the advantage of a structure based approach, using the NCC as an example can be illustrated with the images in Fig. 9 where the same *jpg* encryption method is applied on two different input images. The left pair of images is not recognized by the observers (score=0.892), whereas the right pair is much more easily recognized (score=0.311). Because it is a smooth dark area, the large shadow area under the pontoon leads to an uneven scrambling by the encryption, and thus, the encrypted content still exhibits the shape of the shadow area making it recognizable to a human observer. For both of the lower panel images in Fig. 9, VQIs will estimate that both test images are of poor quality, but won't find any similarities, whereas the NCC might be able to find a matching pattern. Metrics scores are provided beneath the images, however, all four metrics are unable to differentiate both images in terms of recognizability



0.892	Recognition Score	0.311
0.113	LEG	0.085
9.912	PSNR	9.296
0.020	SSIM	0.021
0.015	VIF	0.017
0.120	NCC	0.475

Figure 9: An example of increased recognizability rate due to some particular image properties.

providing highly similar values, whereas the NCC reaches respectively 0.120 and 0.475, thus clearly differentiating the two cases.

Table 4 gives the results of the NCC metric for the entire database and split by test set. Except for two SEnc methods (*jxr* and *265*) the metric performs well, and seems able to better differentiate recognizability scores. The improvements over the standard VQIs is illustrated in Fig. 10, which also shows the poor performance of all metrics on the *265* and *jxr* test sets. The weak performance on the data of these two encryption methods can be explained by the cascaded distortion induced by them, which erases large uniform portions of the image. It is difficult for a correlation based metric to cope with such large smooth image patches. A block-based correlation, followed by a pooling of local scores, might help to improve the predictions in such a framework.

4.2.3. Learning the Recognition Threshold

We do not intend to create a (new) recognition index in this section. Rather, we are interested in whether applying machine learning can improve on the VQIs given above. We directly use the VQI scores on two resolution levels as feature input for a support vector machine (SVM) with a radial basis function. We try to learn a regression model (basically building a new VQI based on the old VQIs). If machine learning can improve the rates above (Table 3) substantially, then this is a clear indication that the features used in classical quality indices

Table 4: Evaluation of structural VQIs on the database, same parameters as in Table 3, for the full database as well as per encryption type.

(a) NCC

Testset	EER [%]	0FNR[%]	$\|\text{MCC}\|$	$\|\text{SROC}_{90}\|$	RMSE	r
all	12.28	100.00	0.574	0.554	0.342	0.695
jpg	5.83	80.00	0.838	0.866	0.138	0.917
jxr	20.83	100.00	0.356	0.164	0.468	0.545
265	42.65	100.00	0.177	0.112	0.607	0.317
j2k	10.02	90.00	0.813	0.418	0.206	0.920
j2kne	5.69	95.65	0.884	0.516	0.197	0.925
fake	6.94	100.00	0.819	0.854	0.126	0.934

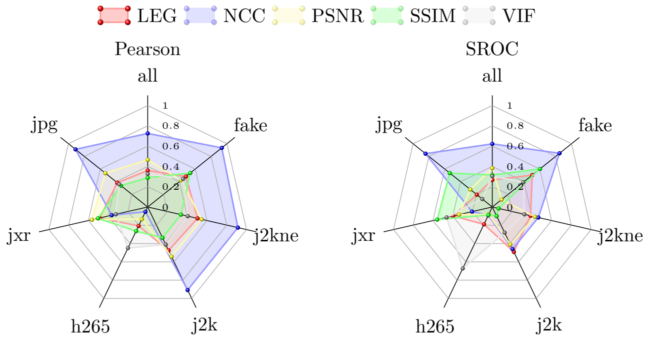


Figure 10: Pearson (left), and Spearman (right) correlations between 5 tested metrics and the observers' recognition rates for each encryption method (from Tables 3 and 4).

can be used for recognition tasks but the valuation of features needs to be reassessed.

We split the database into non-overlapping sets, based on reference images. One set is removed and all others are used for training, then the removed set is evaluated. This is done once per set, in essence building a classification for the whole database by proceeding set by set. The final classification will then be evaluated as given above. The input for the learning will be the output of the LEG, PSNR, SSIM and VIF at two resolution steps, once with full resolution and once downsampled by two. The downsampling should remove some of the noise while keeping larger structures relatively unscathed. As a fitness function we used the 0FNR, EER, MCC and SROC in turn.

The results can be seen in Table 5, split by fitness function ('Target' in the table). From the results it would seem that the regular metric scores are an ill fit for the recognition task. While this combination of VQI scores is certainly better than each VQI individually, the results are still very bad, especially for the typical use-case of 0FNR. The fact that even the SVM has difficulty fitting the VQI score to the recognition data can clearly be seen from examples such as the false positive rate still being at 100% for 0% false negative rate even when optimizing for the 0FNR (jpg, 265). We knew already that the correlation of the VQIs with the recognition data was low (from Table 3), but now we can also see that lifting it to a higher dimension and fitting with an SVM also does not yield much of an improvement.

This indicates that quality and recognition are indeed fundamentally different tasks and that tools which do well on one (quality) are ill suited on the other (recognition).

5. Discussion and Conclusion

As no standardization committee (ITU, VQEG or ETSI) has yet proposed any recommendation to cope with either subjective protocols or objective metrics in the *context of image recognition*, we hereby provide some suggestions and insights on a way to manage a full recognizability assessment (for both the subjective and objective viewpoints). Our recommendations will be structured along three distinct points. At first, we give some comments on the acquisition environment itself, then, we issue recommendations on a possible management of the subjective data and we finally give advice on the proper use of Visual Quality Indices.

5.1. Recommendation for the Acquisition Environment

So far, we have seen little difference between results from the different acquisition environments (and different observer pools). To further test this we can formulate the following Zero-Hypothesis (H_0): There is one underlying (unknown) distribution of observer scores and the results from acquisitions in *CE* and *SE* are drawings from this underlying distribution.

Table 5: Evaluation of the scalable vector regression learning with an SVM based on LEG, PSNR, SSIM and VIF on two resolution scales. The different learning targets are also given.

(a) Scalable vector regression (SVR).

Target	Testset	EER [%]	0FNR[%]	—MCC—	—SROC$_{90}$—	RMSE	r
0FNR	all	9.74	77.78	0.647	0.675	0.168	0.837
0FNR	jpg	9.55	100.00	0.533	0.331	0.145	0.803
0FNR	jxr	5.42	83.33	0.806	0.725	0.142	0.817
0FNR	265	20.38	100.00	0.505	0.564	0.200	0.691
0FNR	j2k	12.38	65.00	0.729	0.627	0.169	0.881
0FNR	j2kne	9.80	60.87	0.793	0.510	0.168	0.889
0FNR	fake	15.83	90.91	0.602	0.540	0.177	0.851
EER	all	8.30	100.00	0.673	0.678	0.259	0.678
EER	jpg	9.55	100.00	0.533	0.349	0.519	0.278
EER	jxr	6.67	100.00	0.759	0.686	0.137	0.849
EER	265	13.24	85.71	0.697	0.708	0.181	0.757
EER	j2k	12.38	85.00	0.703	0.561	0.170	0.878
EER	j2kne	8.83	65.22	0.845	0.465	0.164	0.895
EER	fake	12.91	90.91	0.661	0.532	0.160	0.869
MCC	all	10.11	100.00	0.697	0.627	0.205	0.758
MCC	jpg	9.13	100.00	0.553	0.311	0.140	0.806
MCC	jxr	6.25	100.00	0.723	0.630	0.304	0.644
MCC	265	14.08	85.71	0.532	0.358	0.224	0.631
MCC	j2k	9.95	80.00	0.745	0.667	0.172	0.869
MCC	j2kne	5.69	52.17	0.884	0.497	0.173	0.876
MCC	fake	11.17	90.91	0.666	0.613	0.177	0.832
SROC	all	9.72	100.00	0.632	0.717	0.434	0.527
SROC	jpg	17.64	100.00	0.419	0.381	0.846	0.114
SROC	jxr	5.42	83.33	0.806	0.800	0.176	0.781
SROC	265	14.08	100.00	0.487	0.751	0.533	0.546
SROC	j2k	9.48	95.00	0.743	0.596	0.171	0.886
SROC	j2kne	6.17	73.91	0.845	0.469	0.172	0.893
SROC	fake	8.68	72.73	0.741	0.722	0.201	0.815

The way to evaluate this hypothesis is the 2-sample Kolmogorov-Smirnov (KS) test, [22], resulting in a KS statistics of 0.041 which corresponds to a p-value of 0.54 or 54%. This is clearly not enough to reject H_0.

However, so far we have not found any evidence of difference, which should not be taken as an evidence for the absence of differences.

Still, the conclusion at this point, and without evidence of the contrary, has to be that there is little to no difference between *CE* and *SE*. This would suggest that there is no real benefit of the *CE* over *SE* environment and we therefore **conclude that the use of an uncontrolled environment is acceptable,** if it is needed to keep setup cost and time consumption low.

5.2. On the number of human observers

When running a subjective experiment, besides the protocol itself, one crucial factor that may influence the analysis of the output subjective data is the number of observers. Most reports recommend to enroll at least 15 observers [14, 38]

As explained in [38]: "*The possible number of subjects in a viewing test (...) is from 4 to 40. Four is the absolute minimum for statistical reasons, while there is rarely any point in going beyond 40 (...) In general, at least 15 observers should participate in the experiment.*"

The ITU or VQEG recommendations have been issued for quality assessment tasks not for recognition threshold tracking. The main difference is that for quality evaluation each observer gives a score, while for recognition only a binary decision is recorded. These binary decisions are fused into a final score. This changes the influence of a single observer significantly and likely requires a higher minimum number of observers than recommended for quality assessment.

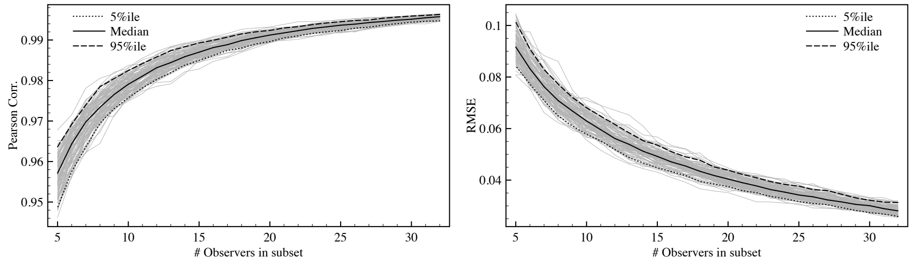


Figure 11: Impact of the size of the panel of observers on the subjective dataset.

In order to test the impact of the number of observers, we consider that the valid observers (after the dendrogram clustering step) represent the ground truth, i.e., a reasonable estimation of the recognizability rate. From this retained set of observers we randomly selected several subsets of observers in order to generate reduced panels of assessors. We chose these reduced panels to comprise between 5 and 32 observers. For each panel size, we randomly generate 100 subsets, i.e., we repeated the test 100 times. For each of the so-obtained subset, we compute both the Pearson correlation and RMSE with the full subjective dataset ($\#O = 67$ for *SPLIT1* and $\#O = 74$ for *SPLIT2*). The result is shown in Figure 11.

As can be seen in this Figure, there is a reasonably small variability among the 100 subsets of observers (100 gray lines). The fewer observers are enrolled, the more the recognizability drifts away from the ground truth. Even though all tested observers are considered *valid*, as given after the clustering stage (Fig. 4), the obtained recognizability rate when the panel is only constituted of 5 observers is quite remote from what we found to be the 'ground truth'.

Clearly the minimum number of 4 observers recommended by the ITU is far too low, and the spread of possible outcomes also varies more strongly. It is difficult to form an authoritative recommendation from this evaluation alone besides: Use as many observers as you can. By keeping in tone with the ITU however we will also **recommend to use *no less* than 15 observers *after* outlier detection**. Optimally, **at least double that number *should* be used**.

5.3. Proper use of Visual Quality Indices

When there is a need to prove either quality (for coding applications), invisibility (in a data hiding scenario), or unrecognizability (in a Selective Encryption framework), most authors turn to Visual Quality Indices that are available in the literature. They apply some state of the art metrics and based on the predicted scores, they eventually compare the performances of their method against some competing techniques. However, the chosen VQI(s) might not be adapted to the tested artefacts, and may not reflect well the human judgement.

The poor overall performance of VQIs only leaves one general conclusion: **VQIs must not be used to evaluate the recognizability of images**. To qualify this statement: This is a general statement. There might well be a VQI which fits a specific encryption method, but this has to be validated and can not be assumed.

This is not surprising since the VQIs were not built with recognizability in mind, and any reasonable performance in this respect would have been incidental. This finding still leaves the community without a proper recognizability indicator. **However, employing the NCC as an example for structural similarity, we have shown that recognition indicators can be designed, but further refinement is still open research**.

Researchers usually opt for using VQIs because of time and monetary cost imposed by the setup of a subjective experiment. We have shown in this work that for a recognizability task, the setup can be significantly lightened, compared to quality assessment or visibility threshold tracking.

If researchers continue to use a VQI for assessing visual recognizability, the onus to prove fitness of the VQI for the task is on them. They might try to ascertain the ability of the VQI to cope with a specific dataset by evaluating it on a limited subset. Indeed, we can not prove there is not a VQI which, for a specific test set, conforms well to human judgment. There is however a pitfall in such a limited subset test as we will briefly outline.

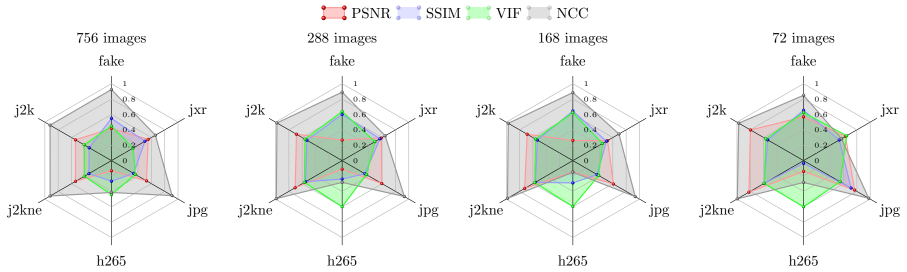


Figure 12: Pearson Correlations between four VQIs and 15 observers for various dataset dimensions

In order to study the feasibility of running a small experiment, and test a few metrics, we have decided to split the whole subjective database into subsets of fewer observers, and most importantly fewer test images. As explained in section 2.1, the full database is composed of 14 images, 6 SEnc methods and 9 SEnc parameters, it is thus composed of 756 test images. Overall 150 observers were enrolled to determine the recognizability of these 756 images (90 under *CE* conditions, and 60 under *SE* viewing conditions). From this database, we have derived 3 smaller subsets:

- 12 images, 6 SEnc methods, 4 SEnc parameters, thus composing a data set with 288 test images.
- 7 images, 6 SEnc methods, 4 SEnc parameters to make a data set with 168 images.
- 3 images, 6 SEnc methods, 4 SEnc parameters, leading to a data set of 72 images.

For each of these reduced data sets, we ran four of the previously tested VQIs. The objective is to ascertain that the best performing metric can be easily determined from a reasonably small subjective data set. Figure 12 shows, as a radar plot, the Pearson correlations between 15 randomly selected observers and four selected VQIs. We present the correlations values for the full size database (756 images) as well as for the 3 reduced versions. It can be observed that among the 4 tested metrics, the VQIs ranking seem to be globally preserved. However, the predicted scores from most VQIs seem to increase when the data set dimensions decreases. In our example this is particularly true for the VIF metric which appears to provide misleading predictions on the smaller data sets for the *265* SEnc method. The lesson learnt from these results is that a subset can be used to evaluate the best VQI out of a set, but the performance on a subset should not be taken as indicative of the performance over a larger set. **The last recommendation therefore has to be that it is possible to use a subset test for VQIs, but the results have to be considered with *extreme* caution**. While the test will correctly show the rankings among the VQIs, i.e., the best VQI for the subset can be found, but the resulting performance is not indicative of the performance over the full set.

5.4. A Dataset for the Development of Recognition Indices

While we were not successful in creating a new recognition index (this was also not the aim of this work), we have shown that objective metrics based on the dataset we collected, and subsequently share with the research community, already show an improvement over existing VQIs. The dataset will be provided to the research community with the purpose of facilitating further research and the development of recognition indices.

Finally, in case of unseen encryption types, we have provided guidelines on the acquisition environment and number of participants as well as the setup of the experiment to facilitate extension of the dataset in a meaningful way.

Acknowledgments

This work was partially supported by the Austrian Science Fund, project no. P27776.

References

[1] Cermak, G., Thorpe, L., Pinson, M., 2009. Test Plan for Evaluation of Video Quality Models for Use with High Definition TV Content. Technical Report. Video Quality Experts Group (VQEG). Video Quality Experts Group (VQEG).

[2] Draschl, M., Hämmerle-Uhl, J., Uhl, A., 2016. Assessment of Efficient Fingerprint Image Protection Principles using different Types of AFIS, in: Proceedings of the 18th International Conference on Information and Communications Security (ICICS'16), pp. 241–253.

[3] International Electrotechnical Commission, 1999. Multimedia systems and equipment - Colour measurement and management - Part 2-1: Colour management - Default RGB colour space. Technical Report. International Electrotechnical Commission. IEC-61966-2-1:1999.

[4] Engel, D., Stütz, T., Uhl, A., 2009. A survey on JPEG2000 encryption. Multimedia Systems 15, 243–270. doi:http://dx.doi.org/10.1007/s00530-008-0150-0.

[5] Grangetto, M., Magli, E., Olmo, G., 2006. Multimedia selective encryption by means of randomized arithmetic coding. IEEE Transactions on Multimedia 8, 905–917.

[6] Hofbauer, H., Autrusseau, F., Uhl, A., 2018. To see or not to see: Determining the recognition threshold of encrypted images, in: Proc. of 7th European Workshop on Visual Information Processing, p. 6.

[7] Hofbauer, H., Uhl, A., 2010. Visual quality indices and low quality images, in: IEEE 2nd European Workshop on Visual Information Processing, pp. 171–176.

[8] Hofbauer, H., Uhl, A., 2011. An effective and efficient visual quality index based on local edge gradients, in: Proc. of the 3rd European Workshop on Visual Information Processing, p. 6pp.

[9] Hofbauer, H., Uhl, A., 2016a. Calculating a boundary for the significance from the equal-error rate, in: Proc. of the 9th IAPR/IEEE Intl. Conference on Biometrics (ICB'16), pp. 1–4.

[10] Hofbauer, H., Uhl, A., 2016b. Identifying deficits of visual security metrics for images. Signal Processing: Image Communication 46, 60 – 75.

[11] Hofbauer, H., Uhl, A., 2018. Applicability of no-reference visual quality indices for visual security assessment, in: Proc. of the 6th ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec 2018), pp. 139–144. doi:10.1145/3206004.3206007.

[12] Hofbauer, H., Uhl, A., Unterweger, A., 2014. Transparent Encryption for HEVC Using Bit-Stream-Based Selective Coefficient Sign Encryption, in: 2014 IEEE Intl. Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 1986–1990.

[13] Huynh-Thu, Q., Ghanbari, M., 2008. Scope of validity of PSNR in image/video quality assessment. Electronics Letters 44, 800–801.

[14] International Telecommunications Union, 2004. Methodology for the subjective assessment of the quality of television pictures. Technical Report. International Telecommunications Union. ITU-R-BT.500-11.

[15] ITU Radiocommunication Assembly, 2002. Methodology for the subjective assessmen of the quality of television pictures. ITU-R BT.500-11.

[16] ITU Radiocommunication Assembly, 2012. Methodology for the subjective assessment of the quality of television pictures. ITU-R BT.500-13.

[17] Jenisch, S., Uhl, A., 2014a. A detailed evaluation of format-compliant encryption methods for JPEG XR-compressed images. EURASIP Journal on Information Security 2014.

[18] Jenisch, S., Uhl, A., 2014b. Visual security evaluation based on SIFT object recognition, in: Iliadis, L., et al. (Eds.), Proceedings of the 10th Artificial Intelligence Applications and Innovations Conference (AIAI 2014), pp. 624–633.

[19] Kauba, C., Mayer, S., Uhl, A., 2016. Image segmentation based visual security evaluation, in: Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec 2016), pp. 1–6. doi:10.1145/2909827.2930806.

[20] Kendall, M.G., 1938. A new measure of rank correlation. Biometrika 30, 81–93.

[21] Lewis, J., 1995. Fast normalized cross-correlation, in: Vision interface, Canadian Image Processing and Pattern Recognition Society, pp. 120–123.

[22] Lopes, R.H.C., Reid, I., Hobson, P.R., 2007. The two-dimensional kolmogorov-smirnov test, in: XI Intl. Workshop on Advanced Computing and Analysis Techniques in Physics Research, p. 12.

[23] Matthews, B., 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. biochimica et biophysica acta (bba) - protein structure 405, 442 – 451. doi:10.1016/0005-2795(75)90109-9.

[24] Ponomarenko, N., Battisti, F., Egizarian, K., Astola, J., Lukin, V., 2009. Metrics performance comparison for color image database, in: Fourth international workshop on video processing and quality metrics for consumer electronics, p. 6 p. URL: http://www.ponomarenko.info/papers/vpqm2009tid.pdf.

[25] Powers, D., 2008. Evaluation: From precision, recall and f-factor to roc, informedness, markedness & correlation. Machine Learning Technology 2.

[26] Preishuber, M., Hütter, T., Katzenbeisser, S., Uhl, A., 2018. Depreciating motivation and empirical security analysis of chaos-based image and video encryption. IEEE Transactions on Information Forensics and Security 13, 2137–2150.

[27] Rieger, M., Hämmerle-Uhl, J., Uhl, A., 2019. Selective JPEG2000 Encryption of Iris Data: Protecting Sample Data vs. Normalised Texture, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'19), pp. 2602–2606. doi:10.1109/ICASSP.2019.8683196.

[28] Sallam, A.I., Faragallah, O.S., El-Rabaie, E.M., 2018. Hevc selective encryption using rc6 block cipher technique. IEEE Trans. on Multimedia 20, 1636–1644. doi:10.1109/TMM.2017.2777470.

[29] Shahid, Z., Puech, W., 2014. Visual protection of HEVC video by selective encryption of CABAC binstrings. IEEE Trans. on Multimedia 16, 24–36. doi:10.1109/TMM.2013.2281029.

[30] Sheikh, H.R., Bovik, A.C., 2006. Image information and visual quality. IEEE Trans. on Image Processing 15, 430–444.

[31] Sheikh, H.R., Wang, Z., Cormack, L., Bovik, A.C., 2005. LIVE image quality assessment database release 2. http://live.ece.utexas.edu/research/quality.

[32] Shekhawat, S., Hofbauer, H., Prommegger, B., Uhl, A., 2020. Efficient fingervein sample image encryption, in: Proceedings of the 8th International Workshop on Biometrics and Forensics (IWBF'20), Porto, Portugal. pp. 1–6.

[33] Spearman, C., 1904. The proof and measurement of association between two things. The American Journal of Psychology 100, 441–471.

[34] Stütz, T., Pankajakshan, V., Autrusseau, F., Uhl, A., Hofbauer, H., 2010. Subjective and objective quality assessment of transparently encrypted JPEG2000 images, in: Proc. of the ACM Multimedia and Security Workshop (MMSEC '10), pp. 247–252.

[35] Stütz, T., Uhl, A., 2007. On efficient transparent JPEG2000 encryption, in: Proc. of ACM Multimedia and Security Workshop, MM-SEC '07, pp. 97–108.

[36] Stütz, T., Uhl, A., 2010. A Survey of H.264 Encryption. Technical Report 2010–10. Dept. of Computer Sciences, University of Salzburg.

[37] Stütz, T., Uhl, A., 2012. A survey of H.264 AVC/SVC encryption. IEEE Trans. on Circuits and Systems for Video Technology 22, 325–339. doi:10.1109/tcsvt.2011.2162290.

[38] Telecommunication Standardization Sector of ITU, 1996. Telephone Transmission Quality audiovisual quality in multimedia services. ITU-T REC P.910.

[39] Union, I.T., 2012. General viewing conditions for subjective assessment of quality of SDTV and HDTV television pictures on flat panel displays BT Series Broadcasting service. Technical Report. International Telecommunications Union. ITU-R BT.2022.

[40] Unterweger, A., Uhl, A., 2012. Length-preserving Bit-stream-based JPEG Encryption, in: MM&Sec'12: Proceedings of the 14th ACM Multimedia and Security Workshop, ACM. pp. 85–89.

[41] VQEG contributors, 2010. Hybrid Perceptual/Bitstream Group TEST PLAN - Draft Version 1.9. Technical Report. Video Quality Experts Group (VQEG).

[42] Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. on Image Processing 13, 600–612.