Cover Page

1) Title of the paper:

LOW QUALITY AND RECOGNITION OF IMAGE CONTENT

2) authors' affiliations and address:

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

LTeN Polytech'Nantes, U6607 and RMeS, INSERM U1229
Rue Ch. Pauc, La Chantrerie, 44306 Nantes, France.

3) e_mail address:

hofbauer@cs.sbg.ac.at

Florent.Autrusseau@univ-nantes.fr

uhl@cosy.sbg.ac.at

4) Conference & Publisher information:

IEEE Trans on Multimedia - TMM
http://www.signalprocessingsociety.org/tmm/
http://www.ieee.org/

5) bibtex entry:

```
@article{IEEE-TMM2021,
  author = {Hofbauer, H. and Autrusseau, F. and Uhl, A.},
  title = {Low Quality and Recognition of Image Content},
  journal = {IEEE Transactions on Multimedia},
  year = {2021}
}
```

Low Quality and Recognition of Image Content

Heinz Hofbauer[1], Florent Autrusseau[2], and Andreas Uhl[1]
[1]Paris Lodron University of Salzburg, Department of Computer Sciences
Email: {hofbauer, uhl}@cs.sbg.ac.at
[2]LTeN, Polytech'Nantes, and Inserm UMR-1229, RMeS, University of Nantes, France
Email: Florent.Autrusseau@univ-nantes.fr

Assessment of visual encryption of video and image content requires a reliable estimation of content recognizability and low quality. As pointed out in the literature, current methods are insufficient and research into this topic, as well as into the relation between low quality and recognizability, is still lacking. This lack of research is primarily due to a lack of data. To improve on the status-quo we have taken a recognizability database and performed a subjective quality evaluation on a subset of the images. This gives us a new database with both subjective recognizability and quality information and allows to delve into the relation between low quality and recognizability. We analyze the relationship between quality and recognizability as well as the predictive quality of state of the art visual quality indices. We show that the visual quality indices are poor indicators for the estimation of recognizability. Furthermore, we show that they must be a poor fit because of the disparity between two distinct perceptual tasks: quality and recognizability.

***Index Terms*—Selective encryption, image recognition, image quality, human visual system, visual quality indices**

I. Introduction

This paper is not about encryption. It is about the relation between recognition, i.e., what is the content of the image, and quality, i.e., how nice does this look. But the reason why we look at the relation between quality and recognizability is very much due to encryption. So we will briefly give an overview of selective encryption and how matters of quality and recognizabiltiy relate to it.

Encryption of image content is an active field of research [1]–[5]. For selective encryption a state-of-the-art cipher is used, e.g., Advanced Encryption Standard (AES), where the security of the encrypted parts is already known and encrypted parts are considered secure. For reasons of speed, usually only a selected part of the data is encrypted. The overall security comes from the data that was selected for encryption, or conversely which data was left in plain-text. Therefore, an analysis of the remaining visual information, which can be extracted from the non-encrypted parts of the data, is necessary.

There is a special case for selective encryption which is called "format compliant encryption". The definition is that a selective encryption scheme is format compliant if a standard compliant decoder can decode the encrypted format without crashing (this is the source of images in the database published with this paper). The benefit of that is that the encrypted data can be used just like regular data. For example, with a careful selection it is possible to use an encrypted video as a low quality preview which can be upgraded to full quality with the key. Depending on the application scenario, the resulting quality can range from "preview quality" to the content should be "unrecognizable".

This is where the notion of quality and recognizability by human observers becomes a primary concern. The typical use cases for selective encryption, providing previews or preventing an enjoyable consumption by human observers (see [6] for typical use cases), directly aim the consumption by humans.

Subjective quality and recognizability assessment is thus needed to ascertain the proper protection by the encryption. To constantly evaluate such systems by actual human observers would quickly become overly time consuming and costly. The obvious alternative is to utilize visual quality indices (VQIs) which are built to emulate the way the human visual system (HVS) assesses the quality. For the development of such VQIs, databases of distorted images with a quality score provided by actual human observers are utilized. There exist plenty of subjective (human assessed) and objective (issued by computer programs) datasets for still, natural image quality assessment [7] or for video quality assessment [8]. Subjective and objective quality assessment studies have also been devoted to Depth Image Based Rendering (DIBR) [9], others have focused on 3D watermarking quality assessment [10]. All these works concern very high quality images evaluation, but only very few works have been conducted on the quality evaluation of selective encryption.

This in turn means that VQIs are trained primarily on high quality databases (because that is what is available) to emulate the human perception of high quality content. It has been shown, [6], [11] that traditional visual quality indices are not well equipped to handle images that move too far away from high quality, and consequently fail at the task of evaluating the content and quality of selectively encrypted data. Indeed, the traditional VQIs are designed in such a way that a special weighting is applied onto some components of the image where the perception is more sensitive. In other words, these quality metrics are tuned to look for differences in a predefined quality range, usually on high quality images to minimize the impact of encoding, and hence are not adapted to handle very strong distortions. Producing new visual quality indices which can handle selective encryption is also a field of active research [12]–[15]. The development of visual quality indices requires either expensive human observer experimentation or a solid database of such observations for evaluation and design.

So the current state of affairs is that the available VQIs are a poor fit for evaluating the low quality of images which are encrypted in a medium to strong fashion. Further, there is little to no work done on the recognizability of strong to very strong selective encryption. Even worse, we do not even know the relationship between quality and recognizability. Still, VQIs (mostly PSNR and SSIM because they are widely known and readily available) are used as a primary tool for evaluating the performances of all encryption strengths, primarily because they are the only tools available.

For the development of visual quality indices for mid- to low-quality images there are, to the best of the authors knowledge, two databases: [16] and [17]. This is not an optimal situation but at least allows for the development on one database and the testing on the other. For the development of indices which can handle content intelligibility the situation is somewhat more complicated as shown in [18] which also introduces the first database with a recognition score for images encrypted on the border of content recognition. The database presented in [17] also contains a recognizability score (denoted content leak information), which is directly rated by observers when comparing the original to the encrypted image, in addition to the visual quality data. The presentation of the original coupled with the tendency of the human visual system to find patterns strongly influences the content leak information. The translation of this content leak information [17] to the recognition index [18] is difficult at best. Specifically, both are susceptible to pareidolia, i.e., the tendency for incorrect perception of a stimulus as an object, but the method in [18] is designed to catch that (multiple originals) while the setup in [17] is not (single original). The data in [17] is still useful for the security evaluation of encrypted images, but not the evaluation of the recognition threshold.

So there is a lack of data which makes it hard to produce quality or recognizability indices for strong to very strong distortions. In addition, we do not know if the VQIs which are available could properly evaluate recognizability, primarily because the relation of quality and recognizability for the HVS in this context is not researched at all. What we do know however, from [18], is that the VQIs most frequently used for this task are not able to perform well in this capacity.

In this paper we present a subset of the database from [18] to a panel of human observers to get an evaluation of perceived quality for encrypted images at the recognition threshold. This is not only a further database for the development of quality based indices for selective encryption but also allows us to study the differences and commonalities between quality and recognizability, a hitherto unstudied subject. What makes this even more interesting is that the general usability of quality indices for the purpose of evaluating recognizability is assumed, that is recognizability is seen as an extension of quality. This is partially a pragmatic use of what is available, i.e., visual quality indices, and partly the reasoning that an unrecognizable image will 'of course' be of low quality, after all, the only content left are distortions. Once started, this practice is repeated, largely without thinking due to precedence of the same practice in literature. To take a closer look at this practice is long overdue and a strong motivation for the evaluations in this paper.

The rest of the paper is structured as follows: The database, how it was collected, its content, and where to get it, is described in Section II; The analysis of the relations between quality and recognizability, i.e., the conformance of quality, perceptibility, and various visual quality and recognition indices, is described in detail in Section III; Some topics are not directly related to the experiments, but are still important to discuss, those are presented in Appendix A; The conclusion, Section IV, gives a recap of our findings and concludes the paper.

II. Database

In this section we will describe the images contained in the database, the setup how the acquisition was performed, the handling of outliers and calculation of the final mean observer score. The database is publicly available to facilitate research.

A. Images Contained in the Database

This database is a subset of the USEE database [18]. A subset was chosen so we could perform the subjective experiments with a large enough number of observers. The primary driver behind the decision was how many images we could handle and the cost and time involved.

The USEE database contains 14 images, 12 color and two grayscale, with 6 encryption types in 9 distinct steps for a total of 770 images (including the originals). We reduced the number of images for the USEE Quality (**USEEQ**) database by removing the grayscale images and two of the encryption types (*fake*, *jpg*). Further, we reduced the number of steps between highest and lowest quality to 6 (from 9 in the USEE database). This means the USEEQ contains a total of 288 images ($12 \times 6 \times 4$). Fig. 1 shows examples of the encryption types for the lighthouse image.

Reproducible Research: The database will be made available at http://wavelab.at/sources/USEEQ. It will contain the subset of images from the USEE[1] database, the individual scores per observer, and the outlier pruned mean observer score (MOS).

B. Encryption Types

We will briefly repeat the encryption types here, more detailed information is contained in the README of the USEE database, as well as the specific parameters used for each image.

H.265: The approach in [19] utilizes encryption of sign coefficients in HEVC data. The encryption is converted to still images by using videos of 1 frame length and applying the encryption to that sequence. Since a single frame encryption has a rather limited impact on quality an iterative method of repeated encoding and encryption cycles was used.

j2k* and *j2kne: The method from [20] encrypts a JPEG2000 file in either layer or resolution progression by encrypting codeblock contributions to packets while maintaining signal markers. The difference between *j2k* and *j2kne* is that regular *j2k* uses error concealment during decoding, which tries to improve the image quality in case “strange” code block content

[1] Available online at: http://wavelab.at/sources/USEE

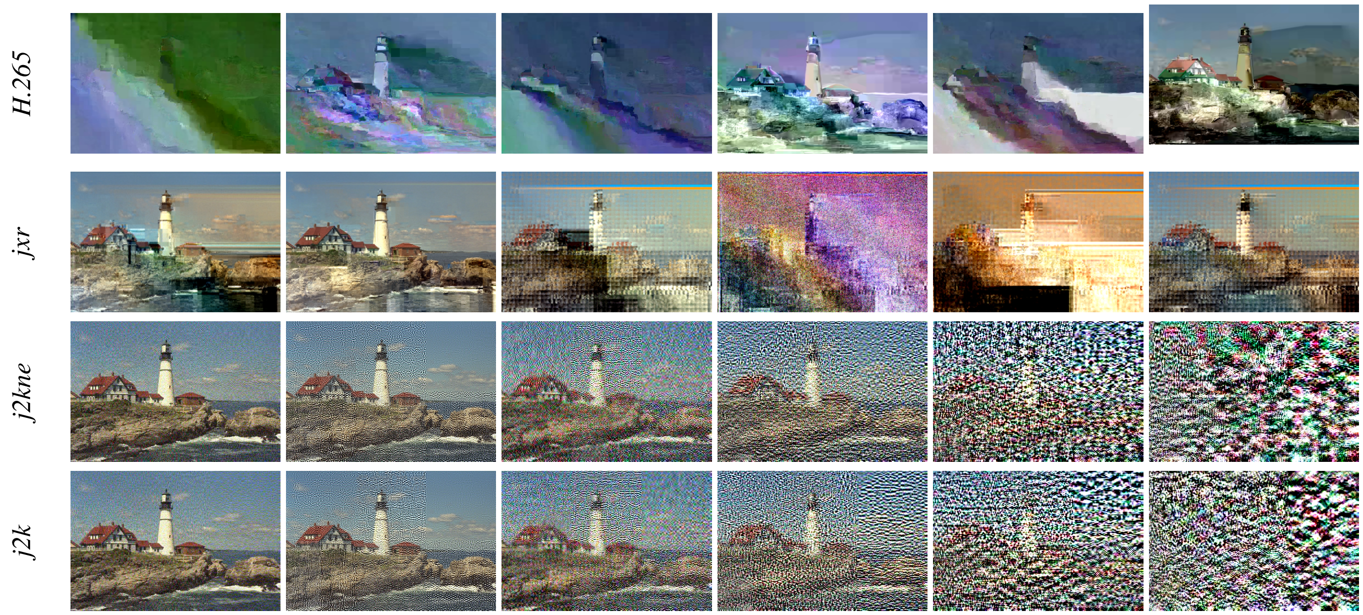


Fig. 1: Example of the encryption steps from the database for the lighthouse image.

arrives (which is the case for encrypted parts), while *j2kne* has turned error concealment off.

jxr: The encryption method for JPEG XR is proposed in [21] and uses 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 and was used to get to desired steps in quality.

C. Acquisition of Human Observer Scores

Forty five observers were enrolled to take part of this study. All observers were either students or staff of the University of Nantes (France). The observers were asked to wear their contact lenses or spectacles during the experiment, they were screened (using Snellen eye chart and Ishihara color plates) to ensure they had a normal acuity and color vision. The observers were paid for their contribution to the experiment, which was completed on average in about 26 minutes per session. A statistical analysis (dendrogram) [22], [23] was conducted in order to detect any possible inconsistent scoring. Of the 45 enrolled observers, based on the deviant subjective scores, 9 observers were discarded in the study. The average age of the remaining 36 observers is 32.94 years old.

The experiment was conducted under standardized viewing conditions. The room illumination was set to 10 Lux, the maximum screen luminance was 200 cd/m^2, the screen gamma function was 2.20 and the contrast ratio/black point of 2 cd/m^2. Our setup was thus compliant with the recommendations by the International Telecommunication Union (ITU) [24]–[26]. The subjective protocol was set in accordance with the ITU recommendations.

The protocol being used here is a “Paired Comparison” setup [27] with a continuous quality scale. Two images were shown side by side on the monitor, the original image was displayed on the left, and the impaired image on the right half of the screen. A grey background was surrounding the images. A horizontal scroll bar was positioned beneath the images, this latter allowed to score the images between 0 and 100. The observers were asked to modulate the scroll bar according to their perception of the distorted image quality as compared with the original. On average, the allocated scores were rather low with an average score of 30.57 and a standard deviation of 5.61.

At this point, thanks to the previously collected recognizability scores [18], and the quality assessment presented here, we have at our disposal, for each image of the USEEQ dataset, two subjective scores: The recognizability Mean Opinion Score (MOS_R) [18] and the quality score (MOS_Q), collected in the experiments as described above.

III. Analysis of the Relation Between Quality and Recognizability

In [6] the authors showed that the usual visual quality indices, which are built for relatively high quality imagery, are not well suited to assess low quality, as in strongly encrypted, images. Furthermore, they pointed out that the lack of a recognizability database prevents any evaluation for the recognition of image content, which is important for confidential encryption. This led to the generation of a recognizability database in [18], which was then used to evaluate visual quality indices to be used as recognition indices. The authors showed that traditional visual quality indices are poor recognition indices. They also proposed using a structure based index (the NCC), in an attempt to create a better index.

Overall, the result from [18] is that the consolidation of recognizability and (low) quality is difficult. However, in [18] no human evaluation of the low quality is available, meaning the assessment of quality and recognizability is only speculative. The relation between high and low quality has to some extent been looked at in [6], but the relation of quality and recognition has not been looked into yet due to a lack of data. We now have a database which has images around the recognition threshold annotated with quality information. This allows us to look into the relation of quality and recognition, with the main goal of understanding why visual quality indices perform so poorly.

As the discussion and analysis can be long, and at times very detailed, we endeavor to succinctly summarize the most relevant results from the longer subsections into an "in brief" paragraph at the end. This allows to skip topics of less interest to the reader without missing the big picture. Please also note that, with the space available, we can not describe all measures and evaluation scores in detail. While we provide a brief description, if a more in-depth description is desired the cite literature should be consulted. Specifically, the reader is referred to [18][2] which contains a more in-depth description of the source database, the encryption types, and the acquisition of the recognitions score; this paper relies heavily on the database and data from that paper.

TABLE I: Experiments on the quality subset

Testset	EER [%]	0FNR[%]	|MCC|	$\|\text{SROC}_{90}\|$	RMSE	r
MOS_Q	4.60	71.43	0.824	0.888	0.562	0.676
VIF	20.62	92.86	0.442	0.287	0.823	0.384
LEG	22.95	100.00	0.504	0.362	0.747	0.335
SSIM	28.42	96.43	0.331	0.182	0.706	0.216
PSNR	17.10	100.00	0.487	0.357	6.325	0.451
NCC	16.91	100.00	0.553	0.593	0.391	0.668
LE	44.05	96.43	0.197	0.040	0.829	0.295
NSD	46.39	100.00	0.137	0.062	0.458	0.035
DBCNN	28.42	100.00	0.312	0.199	0.626	0.223
HyperIQA	24.90	100.00	0.335	0.393	2.739	0.276
NIMA	44.34	100.00	0.162	0.025	1.288	0.058

A. Conformance of Quality and Recognition Scores

The first experiment is to repeat what was done in [18], where only the MOS_R is available, and see how human quality assessment (MOS_Q) relates to recognizability. For details about the reported values see the above cited paper and the papers given in the following brief description. The direct relation between quality and recognizability can be investigated by assessing the MOS_Q and MOS_R scores with the following, well known, measures: the root mean squared error (RMSE) [28], a linear correlation (Pearson's r) [29], and a rank order correlation (Spearman's Rank order correlation (SROC) [30]). In [18] the use of *$SROC_{90}$*, spanning 90% of the recognition scores, is suggested over the full SROC. The reason being that the high number of completely visible images (which have the same rank) messes with the overall rank assessment of the SROC, compare Fig. 2. The removal of 10% is based on an assumed error rate (miss-clicks and such) of users. The images can also be classified into recognizable and unrecognizable classes based on the quality. The classification results in true positives, false positives, false negatives and false positives and based on those occurrences the following statistics can be calculated. Two reported values are operating points on the well known receiver operating characteristic [31]. The equal error rate (EER) is the operating point where the false positive and the false negative rates are equal, it is primarily useful for the comparison of methods and significance calculation [32]. The other is the false positive rate at the threshold where zero false negatives are reported (0FNR) as this is the point of interest (no insecure images are reported as secure) when assessing the security for encryption. Finally, Matthews correlation coefficient (MCC) [33], [34] is related to the chi-squared statistics for the contingency table. It is an especially good fit for classification problems where the class size is unbalanced, as is the case here.

Since USEEQ is a subset of the USEE database we also repeat the VIF [35], LEG [35], SSIM [36], PSNR, and NCC [37] evaluations on this reduced subset to be able to properly compare them. This list is extended by two indices specifically designed for the assessment of selective encryption strength, the local entropy measure (LE) [38] and The neighborhood similarity degree (NSD) [39]. Further, we extend the list of VQIs to include a sampling of recent Convolutional Neural Networks (CNN) based methods, namely DBCNN [40], HyperIQA [41] and NIMA [42]. These were used as an end user would, i.e., no specific training except were required (the last training step of the DBCNN on the LIVE database was performed as given by their instructions). The NIMA is somewhat noteworthy in that it not only used the TID [43] image quality database for training but also the AVA database [44] which is an aesthetic quality database. The results are given in Table I.

The VQIs show the same overall behaviour as in [18], they still are only a weak predictor of recognizability. The same holds true for metrics designed specifically for selective encryption as well as the CNN based methods. The subjective image quality (MOS_Q) on the other hand performs rather well which is kind of surprising. The VQIs are built based on the HVS, although for high quality images, but work poorly on this data while the subjective quality scores (MOS_Q) themselves work decently well. This indicates that there is a difference in how human observers perceive high and low quality data. For example in [17] even for low quality images the content leakage score can be high, meaning the human observer is well capable of differentiating between content and quality. Further, the human visual system is adept at noticing differences of a sufficient magnitude, but that changes with the overall variance in the image (this is usually known as contrast masking). That is, a medium strength error will stand out unpleasantly in a high quality image, while a whole image affected by the same strength error will be rated as a lower quality but no unpleasant error will stand out (this is a topic of research termed just noticeable difference, which is affected by contrast masking, see [45] for an overview). To experience the effect, look at the images in Fig. 1, there is a clear difference in the perceived quality of the *H.265* and *jxr* encryption types, which create interesting color patterns, while the *j2k* and *j2kne* types create a noise like structure which appears to be more unpleasant to view.

From these examples we can see that the human visual system acts very differently depending on the shape and distribution of the noise or distortion. As such, VQIs are trained to resemble the human visual system when assessing high quality data, disagreeing with the human visual system

[2] Openaccess available at http://dx.doi.org/10.1016/j.ins.2020.11.047

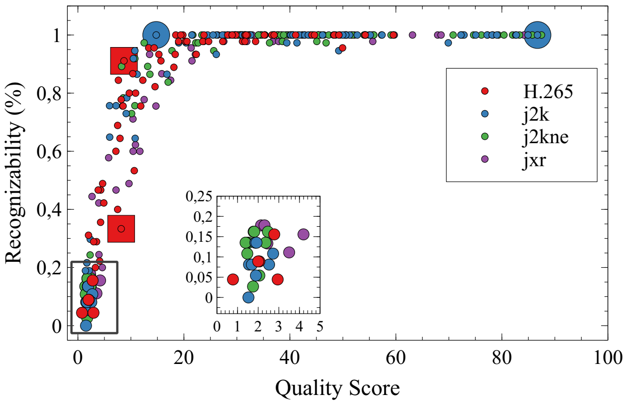


Fig. 2: Scatter-plot of unrecognizability (MOS_R) percentages over quality scores (MOS_Q). Large symbols represents equal quality/recognizability extrema.

when subjected to low quality data. However, the PSNR and the NCC are not based on the human visual system and still perform poorly here.

To take a closer look at this we have plotted the quality (MOS_Q) and recognition (MOS_R) as a scatter plot, shown in Fig. 2. Low (resp. high) recognizability corresponds to low (resp. high) quality. However, the relationship is not linear which can lead to cases where very similar quality scores have highly disparate recognizability scores (and vice versa). Two examples of this (similar recognition and similar quality) are marked in Fig. 2 by the large blue and red symbols. The corresponding images are shown in Fig. 4 and will be discussed later. Even though there is a very distinct non-linear mapping between quality and recognition, they are clearly linked, but not as strongly correlated ($\mathrm{SROC} = 0.859$ and $\mathrm{SROC}_{90} = 0.888$) as expected. For context, the SSIM VQI has a $SROC > 0.9$ on the LIVE database and we expected something similar or better here. This is likely due to the saturation of recognizability (at some point the content is simply recognizable) leading to the following observation.

We also see part of the differences, while quality and recognition are linked the scores overlap only in parts. That is for 90% of the quality range the image is clearly recognizable, while about 90% of the variance in the recognizability is at roughly 10% of the quality range. This becomes clearer when the scores are ordered. Fig. 3 shows both the quality and recognizability scores, once ordered by recognition and once by quality.

What also can be seen from Fig. 3, which is an important point, is that the recognition has a very different range than the quality. At some point the images are recognizable and we have a flat recognizability line, while the quality never bottoms out. This means that users attribute different qualities to images which are basically unrecognizable. This is another strong indicator that the HVS does not only use image content for quality assessment. For unrecognizable images this likely would mean that the aesthetics of the noise, or encryption artefacts, are important for quality, i.e., visually pleasing artefacts lead to a higher quality score than less pleasing artefacts, even if both images are on the same level of recognition.

As an example of the difference in range of similar (by one score) images consider the large symbols in Fig. 2 which represents two cases of this phenomenon. Each pair of images has either a similar quality score (red squares) or a similar recognizability score (blue circles). The corresponding images are shown in Fig. 4. Clearly, quality and recognizability can be highly decorrelated, i.e., for two distinct images a similar perceived quality might result in a significantly different recognizability score (and vice versa). Our method of acquiring recognition scores relies on matching an encrypted image to it's original, to illustrate this we also provide the original as insets in the figure. For a discussion about a potential bias introduced in experiments due to the acquisition method see Appendix A-A.

Fig.3 also shows the separation of the data into high quality, where the images are clearly recognizable, and low quality data, where a distinct drop in quality happens. The range between low and high quality is denoted as medium quality. The quality drop at the border between low and high quality is not very visible in the combined plot but can be easily seen if the plots are separated by encryption type, shown in Fig. 5. The classes low, medium and high quality classes will be used later in the paper (Sections III-E and III-F).

In brief: The quality is an error prone indicator for recognizability of image content. Given that a VQI is itself an error prone predictor of the human visual system we have a doubling up of errors. We also have to keep in mind the different training target for VQIs (high quality) and the effect of contrast masking on the human visual system for bad quality. The result of this is the apparent discrepancy of VQIs and the HVS based MOS_Q as shown in Table I.

This and the attribution of high or low quality to unrecognizable images depending on the visual appearance of errors makes the attempt to directly map quality to recognizability difficult. That is, there is an overlap between recognition and quality but also a large part where there is no proper relation. This can be seen in Figs 2 and 4, where for example about 90% of the quality range is clearly recognizable (related to about 5% of the recognition range). So, while a rank based correlation between quality and recognizability is not likely to succeed a classification into two clusters (recognizable / unrecognizable) based on quality evaluation might work.

B. Visual Quality Indicators as Predictors of Recognizability

A simpler task than a rank correlated estimation of quality is the differentiation between recognizable and non-recognizable images. That is, what is the threshold for a VQI score beyond which images are non-recognizable. This can be simply evaluated by minimizing the total error (E_t) which is the sum of the false non-recognizable rate (E_{FNR}) and false recognizable rate (E_{FRR}) errors which can be defined as:

$$E_{FNR}(T) = \frac{|Q_{NR}(T) \cap R|}{|I|},$$

$$E_{FRR}(T) = \frac{|Q_{R}(T) \cap NR|}{|I|},$$

where I is the set of images, R and $NR = I \setminus R$ are the set of recognizable and non-recognizable images according

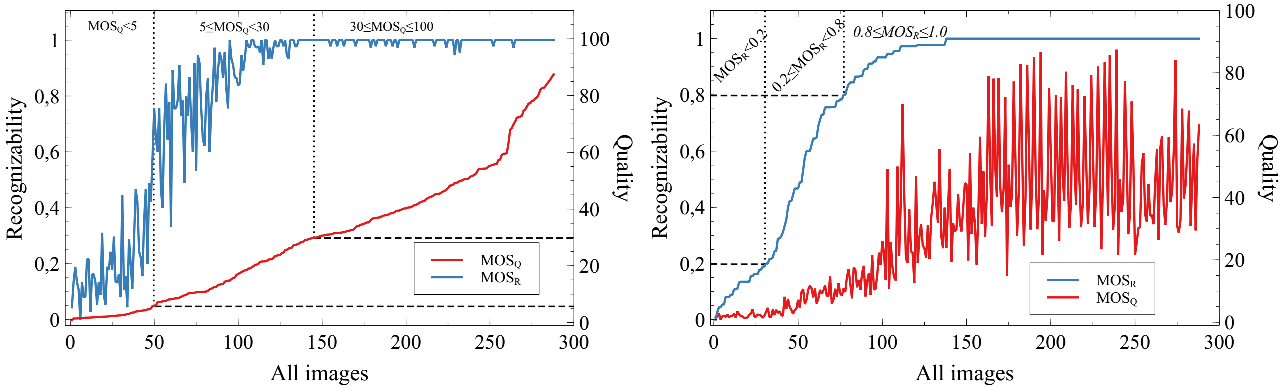
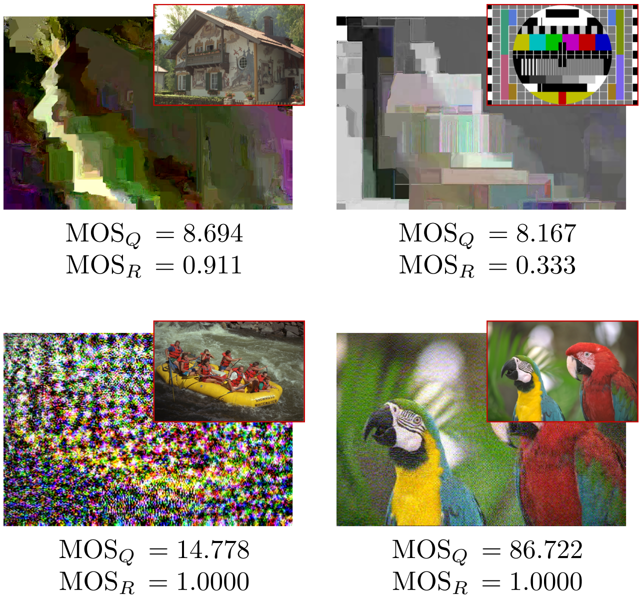


Fig. 3: The figures give the same data, the recognition and quality scores per image, but once ordered by increasing quality (left) and once by increasing recognizability (right).



$\text{MOS}_Q = 8.694$
$\text{MOS}_R = 0.911$

$\text{MOS}_Q = 8.167$
$\text{MOS}_R = 0.333$

$\text{MOS}_Q = 14.778$
$\text{MOS}_R = 1.0000$

$\text{MOS}_Q = 86.722$
$\text{MOS}_R = 1.0000$

Fig. 4: Two examples of quality and recognizability extremas corresponding to the marked pairs in Fig. 2). The original images are shown as insets.

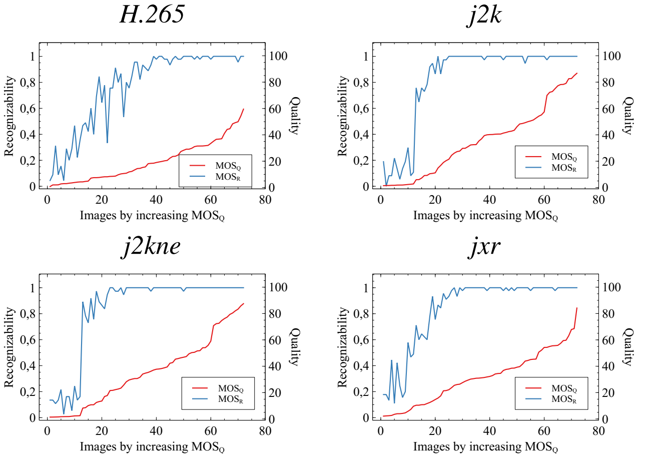


Fig. 5: Ordering of the quality score (monotonically increasing) per encryption method, along with the Recognizability index.

TABLE II: Total error rate ($E_t(T^*_{opt})$ for the optimal threshold (T^*_{opt}) given per VQI and for the quality obtained from human observers (MOS$_Q$).

VQI	T^*_{opt}	$E_t(T^*_{opt})$
MOS$_Q$	0.039	3.47%
NCC	0.032	6.94%
LEG	0.110	8.33%
SSIM	0.010	9.03%
VIF	0.005	9.03%
PSNR	2.650	9.72%
LE	0.003	9.38%
NSD	0.004	9.72%
DBCNN	22.070	9.72%
HyperIQA	19.820	9.72%
NIMA	3.684	9.72%

to human observers, and $Q_R(T)$ and $Q_{NR}(T)$ are the set of recognizable and non-recognizable images according to the quality with threshold T.

The best, as in lowest error, threshold would then be

$$T_{opt} = \underset{T\in[0,1]}{\text{argmin}}\, E_t(T) = \underset{T\in[0,1]}{\text{argmin}}\, (E_{FNR}(T) + E_{FRR}(T))\,.$$

In actuality we have a number of quality values, one per input image, so we simply search over those, i.e. a discretization of the equation. Given that the minimum does not have to happen only on a single value we will use T^*_{opt} as the final (and single) threshold and specify that as

$$T^*_{opt} = \min\, T_{opt}$$

Results are shown in Table II, for the same VQIs we previously used and for the MOS$_Q$ scores obtained from human observers. Clearly quality and recognizability are closely related but not the same, i.e., low but not negligible errors. The same double error as previously discussed for VQIs can be seen here, i.e., VQIs are error prone predictors of the HVS while the HVS is an error prone predictor of the recognition, as seen by the MOS$_Q$ results. In contrast, the NCC, which does not attempt to conform to the human visual system, exhibits a lower error rate than the quality indices. The same can not be said for the PSNR (an objective error metric) as well as the NSD and LE which were explicitly designed for selective encryptions, although the lack of usefulness of the later two beyond the specific encryption they were designed for has already been shown in [6].

Obviously this was calculated on the whole set and is the best result. The question is how well are the different VQIs able to generalize, i.e., what would happen if we applied this to unseen data. The database is composed of four subsets, so we can calculate the threshold for a subset and apply it to the other three sets. This gives us the best result that can be achieved with a given VQI on a single set as well as the generalization performance when applied to a different set. Results are given in Table III. For each VQI the threshold is calculated for one testset, given in *T. Source*, and the total error is given when evaluating classification into recognizable and non-recognizable images of the testset given in the column. Further, each subtable gives the maximum difference per row ($\max \Delta$), representing the generalization properties. Generally speaking the results (outside of the prime diagonal, for which the T^*_{opt} was optimized) falls short of the overall minimum total error given in Table II. Further, the source of the threshold can have a massive impact, compare the rows in Table IIIb where for threshold based on the *jxr* testset the worst (*H.265*) is only worse by 2.78%, but if the threshold is estimated on the *j2k* testset the worst (again *H.265*) has a total error of 52.78%, a degradation of 50%!

Interestingly the PSNR, which does not consider any HVS features, has the best ‘worst case’ of only increasing the $E_t(T^*_{opt})$ by $\approx 6.94\%$. On the other hand the PSNR has the highest total error, see Table II, of $\approx 9.72\%$. This would suggest that the best we can hope for in an unseen data set would be $\approx 16.66\%$ total classification error. The same holds for the NSD but with a slightly worse $\max \Delta$ in one case. The similarity in performance on the recognition task despite a vast difference in quality estimation properties, c.f., Table I, indicates that quality is certainly not the defining factor for recognizability despite what the MOS_Q performance suggests.

In Table II the CNN based VQIs were tied with PSNR in terms of bad performance, but here we see a more diverse set of results. Interestingly the best metric for generalization is the NIMA, which is not based on the visual but rather on the aesthetic quality of images. The HyperIQA and NIMA are relatively consistent, the NIMA closer to the PSNR in terms of performance and the HyperIQA is very close to the LEG, which is the second best traditional VQI in this experiment. So in terms of translating quality to recognizability the CNNs do not outperform the traditional methods. What is however noteworthy is the fact that the NIMA has shown a very poor performance in terms of assessing quality, c.f., Table I, but in terms of deciding between recognizability and non-recognizability is outperforms most other VQIs (traditional and CNN based).

Another important result from this test, and the results from Table IIIa, is that the quality estimation by human observers is also not a good base for the classification of images into recognizable/non-recognizable classes, even though Table II might give that impression.

In brief: We can state two main results: 1) the quality as given by human observers (MOS_Q) is a poor source for the classification of images into recognizable and unrecognizable clusters, and 2) current VQIs are also not well suited to operate this classification (which comes as less of a surprise given that most are built specifically to model the quality estimation by the HVS). 3) CNN based VQIs are roughly equal to the (better) traditional VQIs.

C. Refinement of CNNs and Encryption Differences

An obvious improvement to try is to perform refinement training with the CNNs. For this we perform learning with four folds, selecting one encryption type to exclude and use the other three for training. This again simulates the real world usage where an existing quality index, in this case trained on selective encryption types, is applied to a new encryption scheme (same as the generalization experiments above). As a basis for the refinement we used the pre-trained model provided with the CNNs. The training setup mirrors that of the CNN, basically adapting the provided source code only to the new database but otherwise not changing anything. It should be noted that the DBCNN has two methods, training the full network and training the fully connected layers only, with the later being recommended for refinement learning. Training the fully connected layers only did nothing for the task at hand but training the full model lead to the results given here.

We repeated the experiments from Section III-B, first the Total error over all encryption types given in Table IV.

The main takeaway from this experiment is that training improves overall results (Table IV). However, the CNN based metrics still show a similar performance to traditional VQIs and do not approach the quality of the HVS based quality assessment. On the other hand, the database is limited in size and with a larger database the CNN based VQIs might well outperform traditional VQIs, but this is speculative at this point.

The second takeaway is that the generalization apparently does not improve the performances, in some cases the ’worst case’ got a lot worse than before. So training clearly sacrifices generalization for specificity. This can be seen in the *j2k*/*j2kne* cases where the quality improved when the other was included in the training set as they have very similar distortions. On the flip side, this shows that the distortion types for the various encryption methods (and likely image coding types) are very dissimilar. This has consequences especially for the application to unseen encryption types as the actual performance can not be properly judged.

In brief: Training improves the specificity of the CNNs at the cost of generalization properties. Due to the very different distortion types introduced by the encryption methods generalization to new encryption methods can hardly be judged.

D. Visual Quality Indices as Predictors for Quality on Low Quality Images

In [6], [11] the point was raised that the use of visual quality indices as a recognition index could not be properly evaluated because there is a lack of quality and recognition databases. We will use the evaluation methods from these papers and evaluate the VQIs on the database presented in this paper. This serves a two-fold purpose, 1) we extend the results from the given papers with the missing information and 2) we show (again) that the

TABLE III: Evaluation of generalization properties. Threshold is calculated based on the row entries (T. Source) and applied to the set given in the column. The entries are total error rate, and the maximum difference per row (which have a common source for the threshold) given as an indicator for the generalization potential.

(a) $E_t(T^*_{opt})$ for T^*_{opt} base on MOS_Q.

	Evaluated on				
T. Source	*H.265*	*j2k*	*j2kne*	*jxr*	max Δ
H.265	1.39%	5.56%	2.78%	4.17%	4.17%
j2k	4.17%	5.56%	4.17%	6.94%	2.78%
j2kne	23.61%	9.72%	2.78%	8.33%	20.83%
jxr	4.17%	6.94%	2.78%	2.78%	4.17%

(b) $E_t(T^*_{opt})$ for T^*_{opt} base on **NCC**.

	Evaluated on				
T. Source	*H.265*	*j2k*	*j2kne*	*jxr*	max Δ
H.265	6.94%	11.11%	13.89%	6.94%	6.94%
j2k	52.78%	5.56%	2.78%	19.44%	50.00%
j2kne	11.11%	6.94%	2.78%	8.33%	8.33%
jxr	9.72%	8.33%	8.33%	6.94%	2.78%

(c) $E_t(T^*_{opt})$ for T^*_{opt} base on **LEG**.

	Evaluated on				
T. Source	*H.265*	*j2k*	*j2kne*	*jxr*	max Δ
H.265	5.56%	6.94%	11.11%	9.72%	5.56%
j2k	11.11%	4.17%	9.72%	12.50%	8.33%
j2kne	18.06%	6.94%	6.94%	13.89%	11.11%
jxr	6.94%	11.11%	13.89%	6.94%	6.94%

(d) $E_t(T^*_{opt})$ for T^*_{opt} base on **SSIM**.

	Evaluated on				
T. Source	*H.265*	*j2k*	*j2kne*	*jxr*	max Δ
H.265	5.56%	72.22%	69.44%	22.22%	66.67%
j2k	6.94%	9.72%	13.89%	5.56%	8.33%
j2kne	6.94%	9.72%	13.89%	6.94%	6.94%
jxr	6.94%	11.11%	13.89%	5.56%	8.33%

(e) $E_t(T^*_{opt})$ for T^*_{opt} base on **VIF**.

	Evaluated on				
T. Source	*H.265*	*j2k*	*j2kne*	*jxr*	max Δ
H.265	4.17%	11.11%	13.89%	6.94%	9.72%
j2k	11.11%	11.11%	15.28%	6.94%	8.33%
j2kne	9.72%	11.11%	13.89%	6.94%	6.94%
jxr	13.89%	13.89%	16.67%	4.17%	12.50%

(f) $E_t(T^*_{opt})$ for T^*_{opt} base on **PSNR**.

	Evaluated on				
T. Source	*H.265*	*j2k*	*j2kne*	*jxr*	max Δ
H.265	5.56%	9.72%	12.50%	11.11%	6.94%
j2k	6.94%	9.72%	11.11%	11.11%	4.17%
j2kne	11.11%	13.89%	8.33%	12.50%	5.56%
jxr	6.94%	11.11%	13.89%	6.94%	6.94%

(g) $E_t(T^*_{opt})$ for T^*_{opt} base on **LE**.

	Evaluated on				
T. Source	*H.265*	*j2k*	*j2kne*	*jxr*	max Δ
H.265	5.56%	11.11%	13.89%	6.94%	8.33%
j2k	86.11%	9.72%	12.50%	23.61%	76.39%
j2kne	87.50%	15.28%	12.50%	26.39%	75.00%
jxr	5.56%	11.11%	13.89%	6.94%	8.33%

(h) $E_t(T^*_{opt})$ for T^*_{opt} base on **NSD**.

	Evaluated on				
T. Source	*H.265*	*j2k*	*j2kne*	*jxr*	max Δ
H.265	6.94%	12.50%	13.89%	11.11%	6.94%
j2k	6.94%	11.11%	13.89%	9.72%	6.94%
j2kne	9.72%	12.50%	13.89%	13.89%	4.17%
jxr	6.94%	11.11%	13.89%	6.94%	6.94%

(i) $E_t(T^*_{opt})$ for T^*_{opt} base on **DBCNN**.

	Evaluated on				
T. Source	*H.265*	*j2k*	*j2kne*	*jxr*	max Δ
H.265	6.94%	11.11%	13.89%	12.50%	6.94%
j2k	51.39%	9.72%	15.28%	36.11%	41.67%
j2kne	40.28%	11.11%	13.89%	30.56%	29.17%
jxr	6.94%	11.11%	13.89%	6.94%	6.94%

(j) $E_t(T^*_{opt})$ for T^*_{opt} base on **HyperIQA**.

	Evaluated on				
T. Source	*H.265*	*j2k*	*j2kne*	*jxr*	max Δ
H.265	6.94%	11.11%	15.28%	6.94%	8.33%
j2k	6.94%	11.11%	15.28%	6.94%	8.33%
j2kne	6.94%	11.11%	13.89%	6.94%	6.94%
jxr	6.94%	15.28%	18.06%	6.94%	11.11%

(k) $E_t(T^*_{opt})$ for T^*_{opt} base on **NIMA**.

	Evaluated on				
T. Source	*H.265*	*j2k*	*j2kne*	*jxr*	max Δ
H.265	6.94%	12.50%	15.28%	6.94%	8.33%
j2k	6.94%	11.11%	13.89%	6.94%	6.94%
j2kne	6.94%	12.50%	13.89%	6.94%	6.94%
jxr	8.33%	12.50%	16.67%	6.94%	9.72%

TABLE IV: Total error rate ($E_t(T^*_{opt})$ for the optimal threshold (T^*_{opt}) given per refined CNN (compare of the shelve CNNs in Table II).

VQI	T^*_{opt}	$E_t(T^*_{opt})$
NIMA	2.000	9.720%
DBCNN	5.320	6.940%
hyperIQA	6.150	7.640%

TABLE V: Evaluation of generalization properties for refined CNNs (compare Table III). Threshold is based on row entries (T. Source) and applied to the set given in the column. The entries are total error rate, and the maximum difference per row (which have a common source for the threshold) given as an indicator for the generalization potential.

(a) $E_t(T^*_{opt})$ for T^*_{opt} base on **DBCNN**.

	Evaluated on				
T. Source	*H.265*	*j2k*	*j2kne*	*jxr*	max Δ
H.265	5.56%	20.83%	18.06%	6.94%	15.28%
j2k	6.94%	6.94%	6.94%	6.94%	0.00%
j2kne	6.94%	8.33%	5.56%	6.94%	2.78%
jxr	11.11%	30.56%	27.78%	6.94%	23.61%

(b) $E_t(T^*_{opt})$ for T^*_{opt} base on **HyperIQA**.

	Evaluated on				
T. Source	*H.265*	*j2k*	*j2kne*	*jxr*	max Δ
H.265	4.17%	16.67%	12.50%	13.89%	12.50%
j2k	6.94%	11.11%	6.94%	6.94%	4.17%
j2kne	6.94%	13.89%	4.17%	6.94%	9.72%
jxr	5.56%	11.11%	6.94%	6.94%	5.56%

(c) $E_t(T^*_{opt})$ for T^*_{opt} base on **NIMA**.

	Evaluated on				
T. Source	*H.265*	*j2k*	*j2kne*	*jxr*	max Δ
H.265	6.94%	88.89%	86.11%	6.94%	81.94%
j2k	6.94%	11.11%	13.89%	6.94%	6.94%
j2kne	6.94%	11.11%	13.89%	6.94%	6.94%
jxr	93.06%	88.89%	86.11%	6.94%	86.11%

TABLE VI: The SROC and Confidence (average, standard deviation and signal shape) for the give VQIs based on the quality assessment of the USEEQ database.

		Confidence		
VQI	SROC	μ	σ	Signal Shape
VIF	0.870	0.206	0.145	Bias Low
NCC	0.827	0.638	0.256	Bias High
PSNR	0.641	0.390	0.105	Bias High
LEG	0.624	0.627	0.079	Bias High
SSIM	0.348	0.762	0.183	Stable

correspondence of VQIs to low quality images is low. The specific methods are Spearman rank order correlation (SROC) as well as confidence and signal shape scores which were introduced in [6]. The SROC is simply the evaluation of the monotonous relationship between two scores. The confidence, given as average and standard deviation, gives the possible range of qualities which can lead to an arbitrary but fixed VQI value, specifically this tells us how sure we can be of the quality based on the VQIs output. This is usually not the same over all the quality range, e.g., we can be much more assured of the PSNR when using it on reasonably high quality images than on low quality images, this is given as a signal shape, which can be stable, unstable or biased towards either high or low quality images.

The results for this test are given in Table VI. For the most part the results from [6] are confirmed here, specifically that all VQIs have poor scores for all measures on highly impaired images. As in [6] the VIF is better than the rest, which should not be construed to mean that it is good, there is still ample room for improvement.

In brief: The VQIs, no matter if they are simple statistical measures or relying on advanced HVS models, are very poor indicators of quality on low-quality images.

E. On Using a Fusion of Visual Quality Indicators as Predictors for Quality and Recognition

So far we have looked at VQIs independently to predict recognizability (and low quality scores). There is also the option to combine multiple VQIs. Given that most VQIs utilize different image features to estimate quality, although there is a certain overlap, the combination of all these features might well do what one of them can not. To see the difference in the behavior in VQIs, and also get a different view of the relation between recognizability and quality, we can use Principal Component Analysis (PCA) and biplots, plots of the impact on the principal components per contributing score.

Note: We converted every VQI to a quality index, meaning a high score predicts a high quality. Among the tested VQIs, one (CPA) is actually a distortion measure, meaning a high score predicts a high distortion and thus low quality. This was done so in biplots the closeness of two scores is directly visually obvious, i.e., the angle between the vectors is directly correlated to the influence of the vectors on the subspace projection created by the PCA.

We increased the number of VQIs from previous experiments, by adding the NQM [46], MSSIM [47] and VSNR [48], to increase the number of potential features. Fig. 6 gives the performance of the VQIs for the estimation of quality and recognizability per encryption type. Some metrics do perform quite well when estimating the quality, but most of them fail at recognizability estimation.

A PCA was run independently on each selective encryption method. Fig. 7 shows the biplots, that is the influence on the principal components per input feature, resulting from this PCA. The yellow dots represent the unrecognized images (MOS_R < 0.2), the blue-green dots are the partially recognized images (0.2 $\leq$ MOS_R < 0.8), and finally, the purple dots represent the images being fully recognized MOS_R $\geq$ 0.8). This grouping of recognizability was described in Section III-A and is illustrated in Fig. 3. In the biplots, when two PCA vectors follow the same orientation and have approximately the same length, it means these two variables are strongly connected to each other. Logically, the recognition vector is aligned with the evolution of colored dots. The quality vector (except for *H.265*)

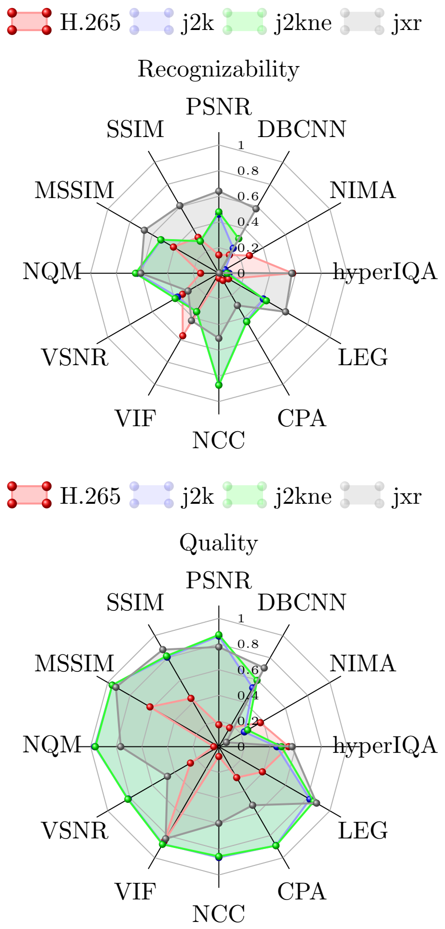


Fig. 6: Comparison of the metrics' behavior against recognizability and quality estimation tasks.

is not completely aligned with the recognizability, expressing a low correlation. This biplot representation is very helpful to determine if one can trust a VQI estimation. That is, if all the VQI variables were aligned, there would be a global agreement on the predicted quality scores, whereas a wide spread, as exhibited here, infers a weak reliability of most metrics.

The first thing to consider is that the recognition and quality are also not aligned very well but also not strongly misaligned, basically reinforcing what we have found previously. The second is that the difference in relation of the VQIs to the recognizability and quality explains the incapability for generalization we have also previously seen. What is more, there is a huge difference in the biplots for each encryption method. Specifically, *j2k* and *j2kne* which are related exhibit a very similar pattern, while the *jxr* and *H.265* are very different. Thirdly, the VQIs are overall more aligned with quality than with the recognizability (which was to be expected given their design target). Finally, some of the VQIs are aligned with each other but the relative alignment is not the same over the different encryption types. This is an indicator that a fusion approach, the use of multiple VQIs instead of a single one, might improve the overall predictive performance.

In brief: The biplots reaffirm that the quality and recognizability are only loosely aligned. The VQIs are aligned better with quality than recognizability, but the relative alignment of the VQIs changes over the testsets. This reaffirms the generalization problem we have seen previously but also suggests that a combination of different VQIs could be beneficial.

F. Predicting Recognizability Classes via Linear Discriminant Analysis

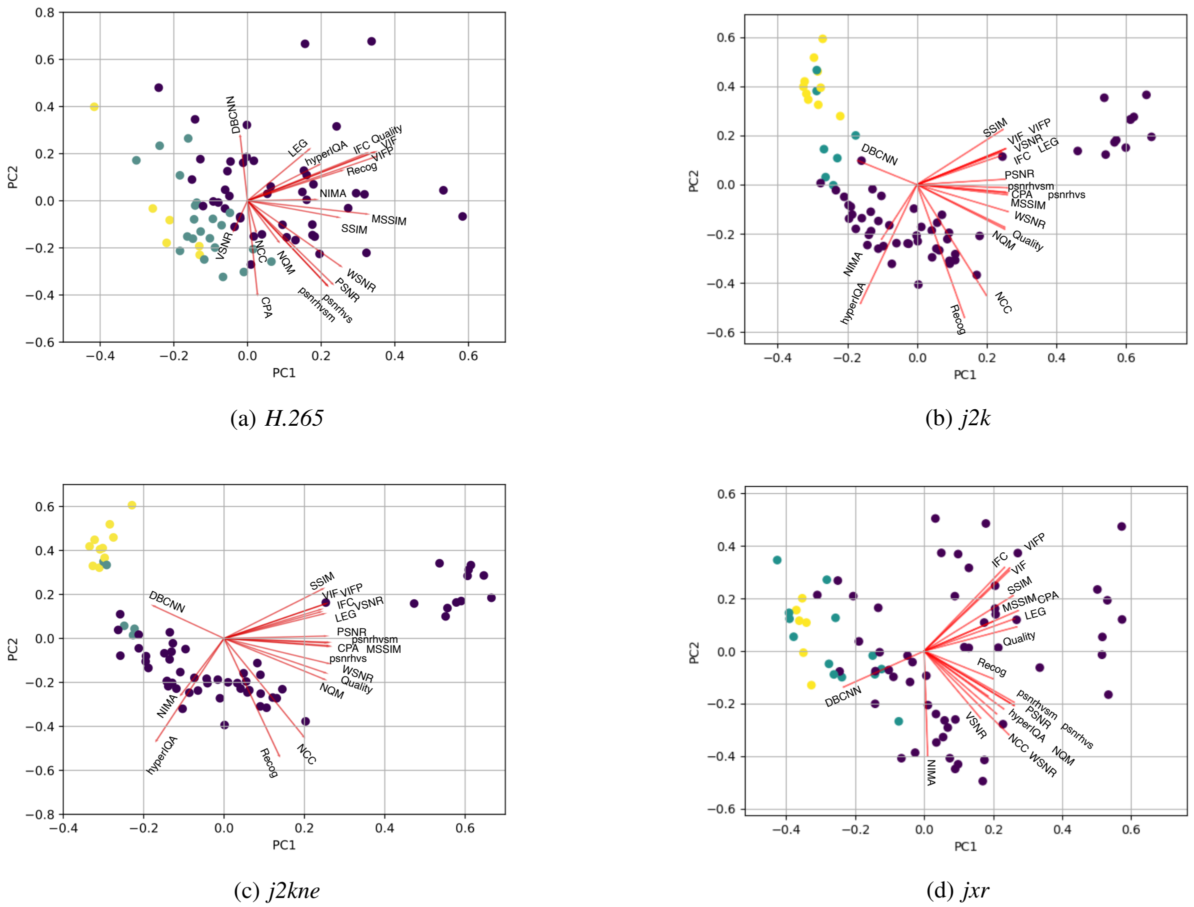
In order to test the potential for a fusion approach we can use a Linear Discriminant Analysis (LDA), similar to the PCA above. The purpose of the PCA is to find the best dimension reduction of our data, i.e., finding linear combinations of the input variables presenting the highest variation in the dataset. The main objective of the LDA is to optimize, for the representation domain provided, the best separation between various classes. The LDA attempts to separate the classes in the best way possible based on input features, in our case the VQI scores per image.

As established before, a ranking for recognizability, or quality, is difficult and not likely to succeed due to the double error problem. To make the prediction more manageable we will split the images into three clusters (see section III-A and Fig. 3): 'Not Recognizable' (NR) images have a recognition score (MOS_R) in $[0, 0.2[$, 'Mostly Recognizable' (MR) are in the range $[0.2, 0.8[$ and 'Fully Recognizable' (FR) in the range $[0.8, 1]$. The goal is to infer the recognizability from quality. In practice we do not have the HVS-based quality and thus will use an ensemble of VQIs instead. We will follow the same basic principle as before in structuring the experiments:

1) Is the basic principle sound? This can also be stated as: are the quality clusters related to the recognition clusters, or can the recognizability be inferred based on quality.
   Note: We have already done this in prior experiments, it is sufficient to look at Figs. 2, 3, and 5 to see that there is a strong relation between the quality and recognizability.
2) Can the clustering be done based on features which are available (VQIs) and how well does that work?
   The target clusters are based on the recognizability as outlined above. And the features are based on the VQIs instead of the HVS-based quality, which we would not have in a practical application.
3) Assuming the clustering based on VQIs works, we have to look at generalization. That is, the available selective encryption types are split into a training and evaluation sets. This allows to simulate the applicability of a trained LDA on an unseen encryption type. Alternatively, train on one and apply to the others. This is much harder of course but can show if the result of LDA based training can generalize, this would reflect what we did in table III.

The results of the LDA, based on the nine VQIs and the MOS_Q, are given in Fig. 8. A relation between recognition and quality is undoubtedly given by the clear separation of the given clusters along the LD1 component. The MOS_Q, which we do not have in practice, is used in the experiment to show that the LDA in principle can handle the prediction.

In Fig. 8 the clustering along LD1 looks promising. To get a better view on the data we split the data by encryption type and repeated the LDA, shown in Fig. 9a. In order to better see



(a) *H.265*

(b) *j2k*

(c) *j2kne*

(d) *jxr*

Fig. 7: Biplots after a principal component analysis applied independently on each distortion.

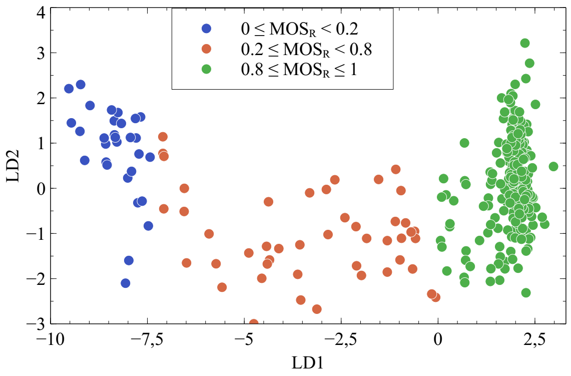


Fig. 8: LDA on the full dataset.

the classification along LD1 we also gave the histograms of a projection of the data onto LD1 in Fig. 9b. The separation of the clusters overall is decent but far from perfect, *H.265* especially has large overlaps. The next step is to repeat the experiments without the MOS_Q and apply the LDA to an ensemble of VQIs.

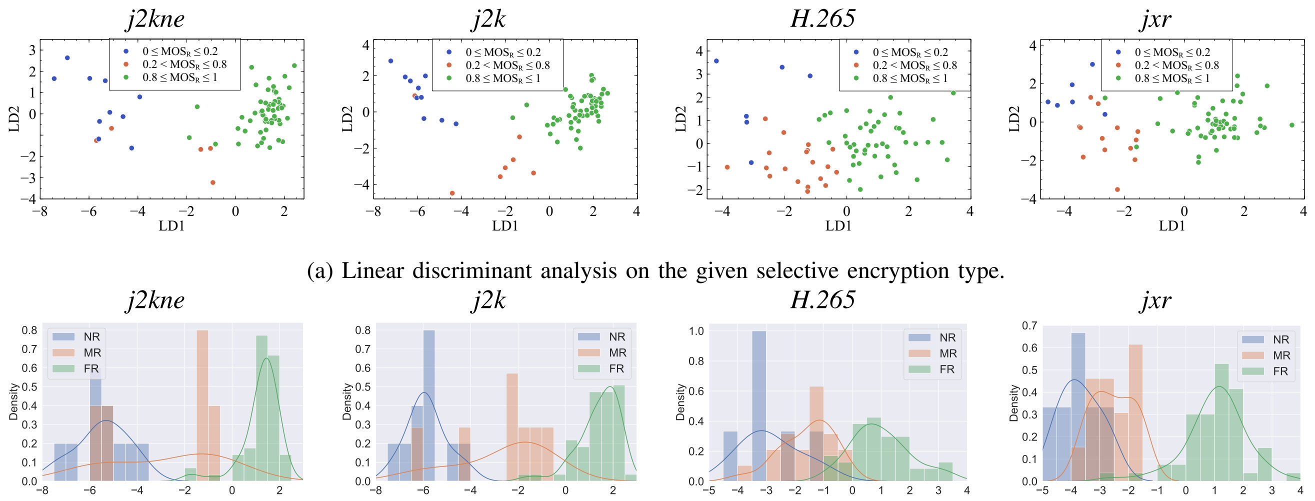
The results are given in Fig. 10a, as before split by encryption type, and the histogram of the projection along LD1 is given in Fig. 10b. The *j2k* and *j2kne* have only minor increase in clustering errors while for *H.265* and *jxr* clusters almost completely break down. The the MOS_Q seems to have only a minor impact on the clustering, but it is only one of many input features in the process. The double error, from VQI to quality to recognition, has a clear impact as can be seen from the lower performance.

Given the weak performance of the LDA when it is based purely on VQIs, Fig. 9, generalization is likely to be weak. And this is exactly the result of the experiments we ran, were we trained on one distortion type and applied the model on other three (the same setup as in III). For reasons of brevity, and due to the totally expected results, we will not show the specific results. As a summary we can say that due to the more limited amount of data for training the overall performance is degraded further. In addition, due to the dissimilarity between the encryption types the generalization was poor.

In brief: The VQIs are not a good source of information to predict the recognizability, even if multiple VQIs are combined in an ensemble.

G. Predicting Quality Classes via Linear Discriminant Analysis

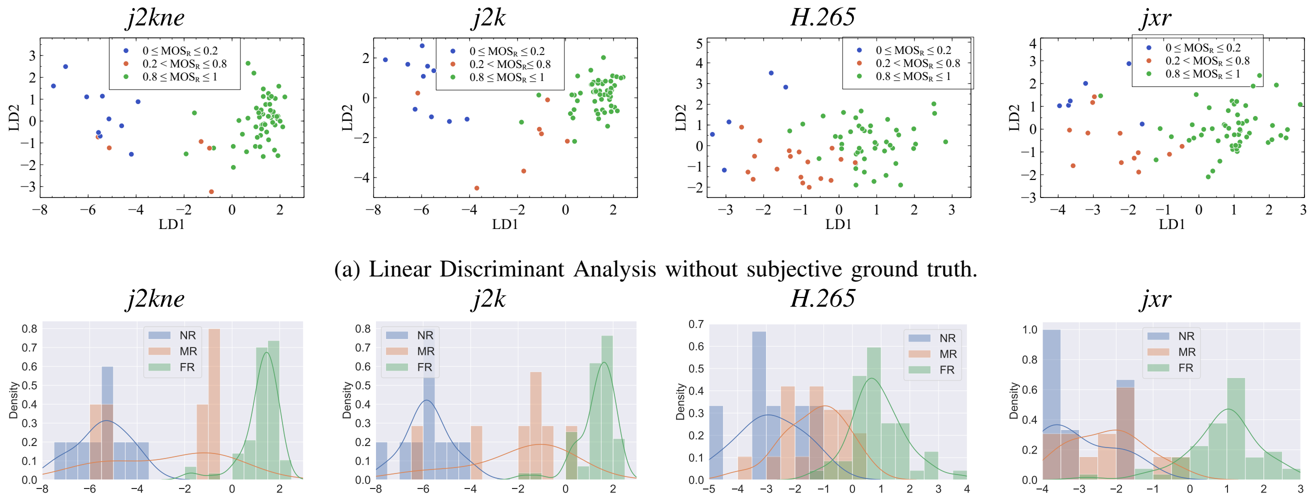
So far we have only looked at the estimation of recognizability from quality. Fig. 11 shows us how the quality based clusters actually match quite well the recognizability of the images. The opposite scenario is depicted in Fig. 12, the very



(a) Linear discriminant analysis on the given selective encryption type.

(b) Histogram of projections of the LDA onto LD1.

Fig. 9: LDA with Recognizability labels, launched on both subjective MOS and VQI predictions.



(a) Linear Discriminant Analysis without subjective ground truth.

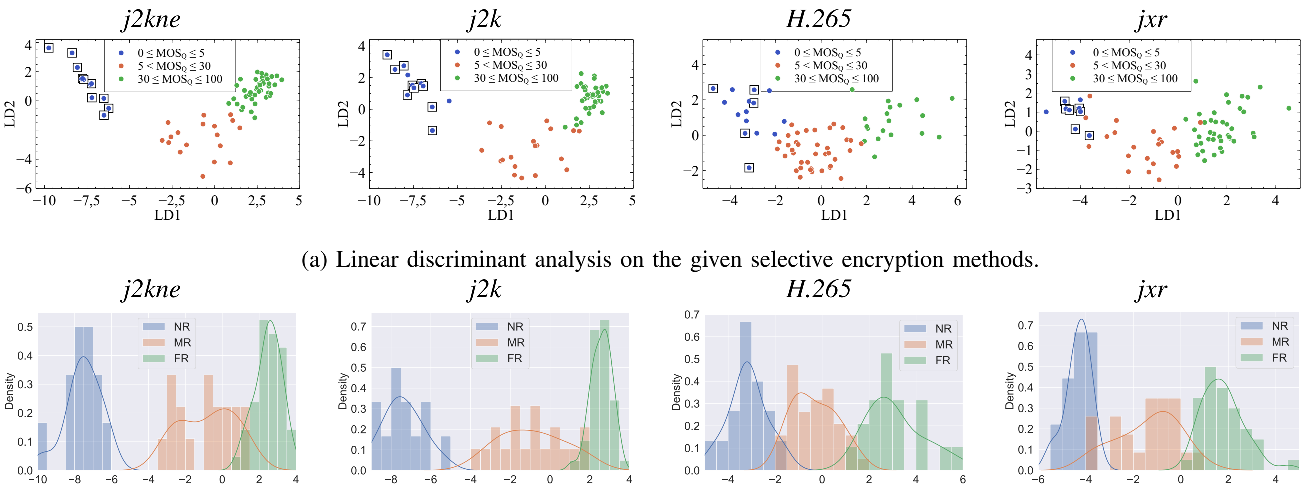
(b) Histogram of projections of the LDA onto LD1.

Fig. 10: LDA with Recognizability labels, launched only on the VQI predictions.

low quality scores (MOS_Q< 5) can be estimated based on the subjective recognizability (MOS_R) and the VQI predictions. In these figures, the LDA was computed using both subjective evaluations (MOS_Q and MOS_R) and VQI predictions. In Fig. 11, the quality scores (MOS_Q) were used along with the predictions, whereas in Fig. 12, the recognizability scores (MOS_R) were used along with the VQI predictions. Once the LDA was computed, we have mapped the unrecognized images onto the quality clusters (Fig. 11) and the low quality images onto the recognizability clusters (Fig. 12). These mapped images are represented as black squares. As can be noticed on these figures, it happens that the LDA seems to successfully cluster together the images having very low MOS_Q and being unrecognized by the observers. However, we can also witness some misclassification issues on Fig.11. Some low quality images (blue dots) were actually recognized by the observers (not surrounded by the black squares). We also can see in Fig.12 that some medium recognizability images (orange dots) actually belong to the lowest quality cluster (surrounded by black squares).

Out of these two scenarios, the first one is the most interesting in practice, as we might expect that some researchers may have launched a subjective quality estimation experiment (based on standardized protocols), but would be in need of estimating the recognizability out of the quality scores. But the reversal of the process should also not be discounted. There are very few VQIs which perform well on the low quality images and basically none that perform well for recognizability. As shown, the development of either would be beneficial as it would also provide an estimation for the other.

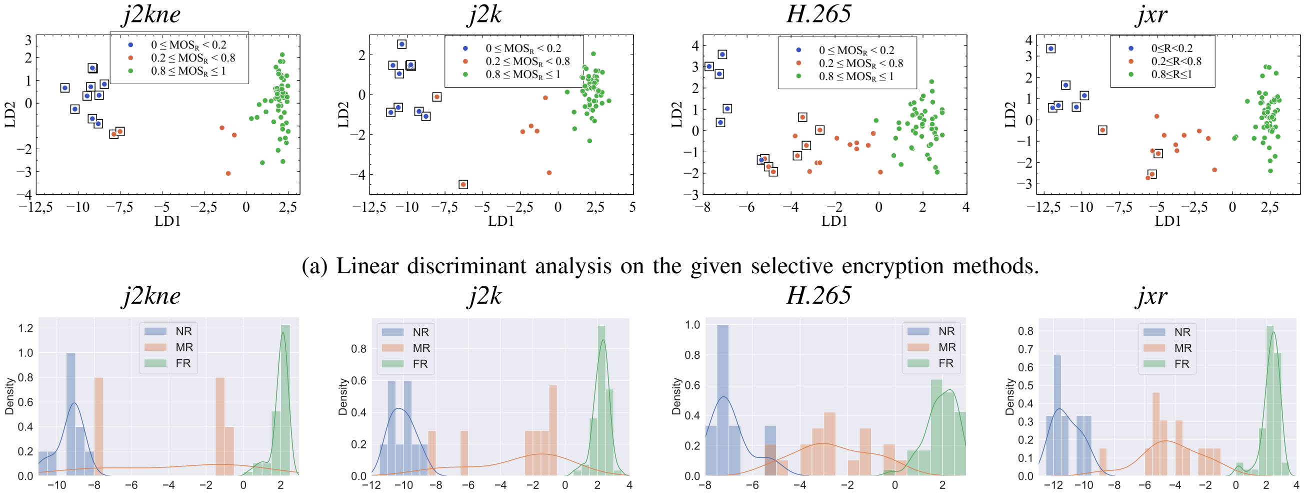
In brief: The unrecognizability and low-quality classes based on MOS_Q and MOS_R have a very large overlap. If a low error estimator for one of those could be produced a somewhat capable estimator for the other would also be available.



(a) Linear discriminant analysis on the given selective encryption methods.

(b) Histogram of projections of the LDA onto LD1.

Fig. 11: Classification of the images into 'High Quality' (HQ), in green, 'Medium Quality' (MQ), in orange, and 'Low Quality' (LQ), in blue. Unrecognized images ($\text{MOS}_R < 0.2$) are depicted by the black square symbols.



(a) Linear discriminant analysis on the given selective encryption methods.

(b) Histogram of projections of the LDA onto LD1.

Fig. 12: Classification of the images into 'High Recognizability' (HR), in green, 'Medium Recognizability' (MR), in orange, and 'No Recognizability' (NR), in blue. Very low quality images ($\text{MOS}_Q < 5$) are depicted by the black square symbols.

IV. Conclusion

We have provided the community with a freely available database of recognition score amended with a quality estimate by human observers. We used this database to look into the relation between quality, recognizability and various visual quality indices.

The overall relation between quality and recognizability can be summed up as "Where quality ends, recognition begins". However, there is a certain overlap. This means that the range of images, where quality scores should be applied versus where recognition scores should be applied, is not clearly separable. This is the unfortunate reality of using the human visual system, which, by default, is subjective and noisy.

We found that the prediction of recognition by using visual quality indices does not yield good results. We have shown that there is a disparity between recognition and quality. We have seen that, in terms of generalization, the PSNR, which is not based on the HVS, beats all the VQIs. However, overall the performance of PSNR (and NCC also) is not good. Further, the disparity between recognition and quality also means that a visual quality index can never properly predict a recognition score.

In the end, the implication of this work is that the automatic evaluation of the recognizability for encrypted images is currently not possible. This means that the use of selective encryption for confidential content is problematic since the correct non-recognizability can not be automatically verified.

Appendix A
Further Discussion

We felt that some points needed addressing without actually impacting the analysis or conclusion of the paper. We compiled them here to have a more streamlined analysis and conclusion.

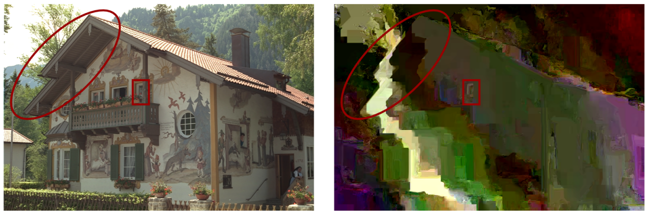


Fig. 13: Some particular shapes within the image that despite a very poor image quality can help the recognizability process.

A. Bias in the Recognition Experiment

By design, we give strong clues to the observers on what the encrypted image contains. In Fig. 13 we show an original image along with one encrypted version. The shape of the roof inside the red ellipse can clearly be matched from within the encrypted image, although it is almost impossible to determine what its content is. Sometimes, an even smaller shape can be recognized, for example the window portion inside the red rectangle can also be identified as long as the observer sees the original image. No quality metric can reasonably give a good score on such a strongly distorted image, leading to a strong double error, see section III.

Thus, the lack of any correspondence between quality and recognizability may actually partly come from the subjective protocol itself. Let us suppose that we designed a completely different subjective protocol, in which, for instance, we show an encrypted image and ask the observers: “What do you see in this image?”. Possible answers might be: a house, a boat, a plane, and so on. Another option would have been to show 2 images side by side, one being the original image and the other one being encrypted. Then, we would ask the following question to the observers: “Are these the same images?”. The recognition rates would likely be different than what they currently are. And who knows, maybe the VQIs would have performed better.

One reason to select the experiment that was chosen is that the recognizability is based on statistical analysis of the results. That means we have to know the chance a random guess is correct as this is the basis of the recognition score calculation. The second version can of course be controlled by showing the correct pair at a fixed rate, e.g. 50%, but then this type of experiment is similar in behaviour to the current protocol.

The experiment where the content of the image has to be described has two problems, one is the chance element. The other is that the labels need to be defined, how many labels, how many per image, etc. This creates a similar problem than for the other experiments in that a certain prior knowledge is available. On the other hand some image content elements might not have a label. For example, in Fig. 13 the window might be recognized, if there is no label for window but for house then the correct label can be guessed even if it can not otherwise be inferred from the image. This setup is therefore quite complicated regarding labeling, what to label, how many labels etc., has to be carefully chosen. In addition, the time investment to evaluate a single image is higher, since the correct labels have to be selected, maybe more than once, and likely from a list which prompts a reevaluation of what is seen in context of the label. This increases the time spent per image, taking observer fatigue into account this necessitates a larger number of sessions and makes the scheduling more complex. The additional time required also comes at additional cost.

Another approach is the one taken in [17] where an original was compared directly to the encrypted version and a subject is supposed to judge the amount of information left in the encrypted image. This is a direct translation of the quality estimation protocol proposed by the ITU to the recognition task. As already discussed this suffers from pareidolia, which makes the proper estimation of a recognition threshold unlikely. Compared to the data from [18] it is, however, more likely to have a better representation of the intermediary part between recognizable and unrecognizable as the quantification is more fine-grained, while in [18] it is binary (recognizable or not).

Concluding, it is (almost?) impossible to design a protocol which gives the required information in a timely manner without introducing a bias. The only way to deal with this is to try and keep the bias small and be mindful of it during the evaluation.

B. Quality and Prediction in the HVS

The HVS can switch between recognition, i.e., what is in the image, and quality, i.e., how nice does this look. But these are distinct processes. There are a lot of images which are beyond quality, that is images which are of such low quality that if you look at them the question “how good does it look” does not arise. The question is rather “What am I looking at?”. This is where quality ends, but there still might be something recognizable. Same on the other end of the spectrum where the image is perfectly recognizable, and therefore the question “What am I looking at?” does not even come to mind. But there is still a quality we can assess.

So the ranges of recognition and quality overlap, but quality extends beyond recognition (on recognizable images) and recognition extends beyond quality (on very low quality images). Since there is an overlap we can predict the quality from recognizability and vice versa to a degree, i.e., the region where they overlap, but not on the whole range because the overlap is only partial.

Due to this overlap, predictions work, but not very well. But because they do work to an extent, it is easy to assume there must be some method which properly predicts one from the other. But because the recognition and quality are not totally aligned in range there might well be no ’perfect’ prediction from one to the other.

The conclusion seems to be that there can be no proper prediction between quality and recognizability! This does not preclude image features chosen specifically for the prediction of one type, e.g., every VQI, from potentially predicting recognizability. The features are image features and not quality features per se, but they were specifically chosen for their alignment with quality, so might not actually be the best choice.

C. Reliance on Visual Quality Indices

It should be noted that we used the visual quality indices to attempt to predict the recognizability purely due to availability. We have already discussed how the double prediction, from

VQI to quality and from quality to recognizability, introduces additional errors. But the VQIs are diverse and readily available and thus are a natural choice to use as sources for the prediction. However, in the long term if a proper recognizability score is required it would be better to cut out the middleman and attempt to develop a recognizability score directly based on the recognizability databases. However, as this paper, and prior papers [17], [18], have shown, the features used in VQIs are not a good fit for the recognizability task.

The development of recognizability score therefore will need a careful analysis of image features which can be utilized. But with the database presented here, and in [17], at least the tools for such a development effort are available.

D. The particular case of the Linear Discriminant Analysis

All throughout this project, several attempts have been made to link the subjective recognizability with both the objective and subjective quality. Various tools have been tested. We have for instance tried to use some feature detection algorithms, such as SIFT [49] or SURF [50]. We also made an attempt to adapt the regular quality VQIs into the specificity of selectively encrypted images. Effectively, we quite often witness a loss of the image's highest frequencies, we have thus decomposed the images into various frequency bands (either in the Fourier domain or in the wavelet domain), and ran independently the VQI within each frequency range. Weighting some of the VQIs (the non HVS-based ones) with the Contrast Sensitivity Function (CSF) [51] has also been considered, in order to give more importance to the lowest frequency components. We have also modified the CSF to give even more weights to the lowest frequencies. Unfortunately, none of these numerous attempts proved to efficiently link recognizability and quality.

The only method presenting interesting results was actually the multivariate analysis which we have presented in this paper. The reason for that is probably because, by design, the PCA or LDA are able to benefit from heterogeneous data, discard irrelevant inputs, while enhancing the relevant ones. As already briefly mentioned in section III-E, each VQI spans a certain range of perceived qualities. Among the tested VQIs, one for instance has been specifically designed and tuned for optimized performance against data hiding scenarios (the CPA [52]). And indeed, the CPA exhibits a better performance than other metrics near the visibility threshold. One is able to express some quality enhancement (VIF), i.e., when a test image has a better quality than the original, whereas the purely statistical VQIs (PSNR or SSIM) are commonly more adequate in a medium quality range. On the other hand, some other VQIs might present better performances within lower quality ranges. By feeding the LDA with predicted scores having such a disparity, the LDA weights will automatically adjust to the metrics being the more relevant to the task at hand. This may explain why the LDA was better able to infer the recognizability based on the objective quality assessment from several disparate predictions.

Finally, an interesting asset of the LDA is its ability to process some completely unrelated data and still produce a succeeding clustering. In our example, the LDA mixes altogether some similarity metrics (either HVS based or purely based on the image statistics) and a distortion index (CPA), along with a structural similarity measure (NCC). We could even imagine blending in some feature detection outputs (such as SIFT or SURF mentioned earlier). No matter the relevance of each of these measures to the task at hand, the LDA adjusts its weights and considers all these diverse measures at its disposal.

TABLE VII: Experiments on the USEE database with refined (on the recognition score) CNNs.

Testset	EER [%]	0FNR[%]	$\lvert$MCC$\rvert$	$\lvert$SROC$_{90}\rvert$	RMSE	r
NCC	13.85	83.33	0.744	0.627	0.294	0.725
VIF	18.51	95.83	0.460	0.304	0.834	0.326
LEG	23.25	100.00	0.368	0.268	0.768	0.356
NIMA	45.54	100.00	0.124	0.073	1.276	0.099
NIMA refined	16.56	97.22	0.515	0.451	6.885	0.682
DBCNN	29.17	100.00	0.284	0.156	0.622	0.265
DBCNN refined	20.08	97.22	0.426	0.422	0.299	0.542
HyperIQA	41.48	100.00	0.184	0.216	2.834	0.108
HyperIQA refined	22.19	100.00	0.587	0.431	0.242	0.679

E. Training CNNs for Recognition Estimation

In the main part we trained on quality as the prediction from quality to recognizability is the primary focus of this paper. However, the quest for a proper recognition measure is still not over, the one prior attempt in the form of the NCC has only been a middling success. The CNNs are well known to be adaptable and training them on the recognition score directly seems reasonable. This is what we did, a refinement training on the whole USEE database [18], with a leave one-encryption out 6-fold training. The same evaluations as in Section III-A were performed and are shown in Table VII. Score for some of the other VQIs as well as the CNNs without refinement training are given for context.

Training clearly improves the performance of CNNs a lot. They are still worse than the NCC, and overall can not be considered good recognition metrics. However, it is unclear if this is the full potential of the CNNs due to the rather limited number of images in the database compared to the usual amount required for CNN training, although this is mitigated somewhat by using a pre-trained model for a warm start. Another note, in contrast to Section III-C where refinement on the DBCNN fully connected layers only did nothing we have the reverse situation here, a full model refinement actually reduced the performance while the full connected layer only refinement lead to the shown results.

Acknowledgment

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

References

[1] S. Kotel, F. Sbiaa, M. Zeghid, M. Machhout, A. Baganne, and R. Tourki, "Efficient hybrid encryption system based on block cipher and chaos generator," in *2016 IEEE International Conference on Computer and Information Technology (CIT)*, 2016, pp. 375–382.

[2] A. I. Sallam, O. S. Faragallah, and E. M. El-Rabaie, "HEVC selective encryption using RC6 block cipher technique," *IEEE Transactions on Multimedia*, vol. 20, no. 7, pp. 1636–1644, July 2018.

[3] M. K. Abdmouleh, A. Khalfallah, and M. S. Bouhlel, “A novel selective encryption DWT-based algorithm for medical images,” in *2017 14th International Conference on Computer Graphics, Imaging and Visualization*, 2017, pp. 79–84.

[4] F. Peng, X. Zhang, Z. Lin, and M. Long, “A tunable selective encryption scheme for H.265/HEVC based on chroma IPM and coefficient scrambling,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 8, pp. 2765–2780, 2020.

[5] J. He, S. Huang, S. Tang, and J. Huang, “JPEG image encryption with improved format compatibility and file size preservation,” *IEEE Transactions on Multimedia*, vol. 20, no. 10, pp. 2645–2658, 2018.

[6] H. Hofbauer and A. Uhl, “Identifying deficits of visual security metrics for images,” *Signal Processing: Image Communication*, vol. 46, pp. 60 – 75, 2016.

[7] C. Yang, X. Zhang, P. An, L. Shen, and C. . J. Kuo, “Blind image quality assessment based on multi-scale KLT,” *IEEE Transactions on Multimedia*, pp. 1–1, 2020.

[8] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, “Objective video quality assessment methods: A classification, review, and performance comparison,” *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 165–182, 2011.

[9] S. Tian, L. Zhang, L. Morin, and O. Déforges, “A benchmark of DIBR synthesized view quality assessment metrics on a new database for immersive media applications,” *IEEE Transactions on Multimedia*, vol. 21, no. 5, pp. 1235–1247, 2019.

[10] M. Corsini, E. D. Gelasca, T. Ebrahimi, and M. Barni, “Watermarked 3-d mesh quality assessment,” *IEEE Transactions on Multimedia*, vol. 9, no. 2, pp. 247–256, 2007.

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

[12] A. S. Abraham, L. R. Nair, and M. S. Deepa, “A novel method for evaluation of visual security of images,” in *2017 International Conference on Networks Advances in Computational Technologies (NetACT)*, 2017, pp. 387–391.

[13] T. Xiang, S. Guo, and X. Li, “Perceptual visual security index based on edge and texture similarities,” *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 5, pp. 951–963, 2016.

[14] T. Xiang, Y. Yang, H. Liu, and S. Guo, “Visual security evaluation of perceptually encrypted images based on image importance,” *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2019.

[15] G. Yue, C. Hou, K. Gu, T. Zhou, and H. Liu, “No-reference quality evaluator of transparently encrypted images,” *IEEE Transactions on Multimedia*, vol. 21, no. 9, pp. 2184–2194, 2019.

[16] T. Stütz, V. Pankajakshan, F. Autrusseau, A. Uhl, and H. Hofbauer, “Subjective and objective quality assessment of transparently encrypted JPEG2000 images,” in *Proceedings of the ACM Multimedia and Security Workshop (MMSEC ’10)*. Rome, Italy: ACM, Sep. 2010, pp. 247–252.

[17] S. Guo, T. Xiang, X. Li, and Y. Yang, “PEID: A perceptually encrypted image database for visual security evaluation,” *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1151–1163, 2020.

[18] H. Hofbauer, F. Autrusseau, and A. Uhl, “To recognize or not to recognize — a database of encrypted images with subjective recognition ground truth,” *Information Sciences*, no. 551, pp. 128–145, 2020.

[19] H. Hofbauer, A. Uhl, and A. Unterweger, “Transparent encryption for HEVC using bit-stream-based selective coefficient sign encryption,” in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Florence, Italy: IEEE, May 2014, pp. 1986–1990.

[20] T. Stütz and A. Uhl, “On efficient transparent JPEG2000 encryption,” in *Proceedings of ACM Multimedia and Security Workshop, MMSEC ’07*. New York, NY, USA: ACM, Sep. 2007, pp. 97–108.

[21] S. Jenisch and A. Uhl, “A detailed evaluation of format-compliant encryption methods for JPEG XR-compressed images,” *EURASIP Journal on Information Security*, vol. 2014, no. 6, 2014.

[22] M. H. Chehreghani and M. H. Chehreghani, “Learning representations from dendrograms,” *Machine Learning*, vol. 109, no. 9, 2020.

[23] H. Hofbauer, F. Autrusseau, and A. Uhl, “To see or not to see: Determining the recognition threshold of encrypted images,” in *Proceedings of 7th European Workshop on Visual Information Processing (EUVIP’18)*, 2018, p. 6.

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

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

[26] ——, “Methodology for the subjective assessment of the quality of television pictures,” 2012, ITU-R BT.500-13.

[27] J. Lee, “On designing paired comparison experiments for subjective multimedia quality assessment,” *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 564–571, 2014.

[28] VQEG contributors, “Hybrid perceptual/bitstream group test plan - draft version 1.9,” Video Quality Experts Group (VQEG), Tech. Rep., 2010.

[29] K. Pearson, “Note on Regression and Inheritance in the Case of Two Parents,” *Proceedings of the Royal Society of London Series I*, vol. 58, pp. 240–242, Jan. 1895.

[30] C. Spearman, “The proof and measurement of association between two things,” *The American Journal of Psychology*, vol. 100, no. 3/4, pp. 441–471, 1904.

[31] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861 – 874, 2006.

[32] H. Hofbauer and A. Uhl, “Calculating a boundary for the significance from the equal-error rate,” in *Proceedings of the 9th IAPR/IEEE International Conference on Biometrics (ICB’16)*, 2016, pp. 1–4.

[33] B. Matthews, “Comparison of the predicted and observed secondary structure of t4 phage lysozyme,” *biochimica et biophysica acta (bba) - protein structure*, vol. 405, no. 2, pp. 442 – 451, 1975.

[34] D. Powers, “Evaluation: From precision, recall and F-factor to ROC, informedness, markedness & correlation,” *Machine Learning Technology*, vol. 2, 01 2008.

[35] H. S. A. Bovik, “Image information and visual quality,” *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[36] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

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

[38] J. Sun, Z. Xu, J. Liu, and Y. Yao, “An objective visual security assessment for cipher-images based on local entropy,” *Multimedia Tools and Applications*, Mar. 2010.

[39] Y. Yao, Z. Xu, and J. Sun, “Visual security assessment for cipher-images based on neighborhood similarity,” *Informatica*, vol. 33, pp. 69–76, 2009.

[40] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, “Blind image quality assessment using a deep bilinear convolutional neural network,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 36–47, 2020.

[41] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. Zhang, “Blindly assess image quality in the wild guided by a self-adaptive hyper network,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[42] H. Talebi and P. Milanfar, “NIMA: neural image assessment,” *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3998–4011, 2018.

[43] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. J. Kuo, “Color image database TID2013: Peculiarities and preliminary results,” in *Proceedings of 4th Europian Workshop on Visual Information Processing (EUVIP’13)*, 2013, pp. 106–111.

[44] N. Murray, L. Marchesotti, and F. Perronnin, “AVA: a large-scale database for aesthetic visual analysis,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2408–2415.

[45] J. Wu, G. Shi, and W. Lin, “Survey of visual just noticeable difference estimation,” *Frontiers of Computer Science*, vol. 13, no. 1, pp. 4–15, 2019.

[46] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, “Image quality assessment based on a degradation model,” *IEEE Transactions on Image Processing*, vol. 9, no. 4, pp. 636–650, Apr. 2000.

[47] Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multi-scale structural similarity for image quality assessment,” in *Proc. 37th IEEE Asilomar Conference on Signals, Systems and Computers*, 2003, pp. 1398–1402.

[48] D. Chandler and S. Hemami, “VSNR: A wavelet-based visual signal-to-noise ratio for natural images,” *IEEE Transactions on Image Processing*, vol. 16, no. 9, pp. 2284–2298, Sep. 2007.

[49] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[50] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, “Speeded-up robust features (SURF),” *Comput. Vis. Image Underst.*, vol. 110, no. 3, p. 346–359, Jun. 2008.

[51] P. Barten, *Contrast sensitivity of the human eye and its effect on image quality*. SPIE Press, 1999.

[52] M. Carosi, V. Pankajakshan, and F. Autrusseau, “Toward a simplified perceptual quality metric for watermarking applications,” in *Proceedings of the SPIE conference on Electronic Imaging*, vol. 7542, 2010.