Signal Processing: Image Communication 46 (2016) 60–75

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

journal homepage: www.elsevier.com/locate/image

Identifying deficits of visual security metrics for images

Heinz Hofbauer*, Andreas Uhl

University of Salzburg, Department of Computer Sciences, Austria

ARTICLE INFO

Article history:
Received 26 August 2015
Received in revised form
10 March 2016
Accepted 3 May 2016
Available online 4 May 2016

Keywords:
Confidence
Image metrics
Partial encryption
Security metrics
Selective encryption
Sufficient encryption
Transparent encryption

ABSTRACT

Visual security metrics are deterministic measures with the (claimed) ability to assess whether an encryption method for visual data does achieve its defined goal. These metrics are usually developed together with a particular encryption method in order to provide an evaluation of said method based on its visual output. However, visual security metrics themselves are rarely evaluated and the claim to perform as a visual security metric is not tied to the specific encryption method for which they were developed. In this paper, we introduce a methodology for assessing the performance of security metrics based on common media encryption scenarios. We systematically evaluate visual security metrics proposed in the literature, along with conventional image metrics which are frequently used for the same task. We show that they are generally not suitable to perform their claimed task.

1. Introduction

The claim of visual security metrics (security metrics for brevity) is usually the ability to assess the functionality of an encryption method based on the output of the encryption of visual data. In particular, the evaluation of an encryption method is only based on the visual output (i.e., the ciphertext), which is either an image or video. While such metrics are often created in conjunction with a specific encryption method and tested, if at all, only for this encryption method, the claim to perform as a security metric is usually universal. Furthermore, regular image quality metrics, such as the frequently used PSNR and SSIM, are also utilized in the literature to evaluate encryption methods [1–4].

The problem with the evaluation of security metrics is the fact that there is no established testing methodology. Thus, even if security metrics are tested, the test is usually based on the evaluation procedures for regular image metrics, which are not sufficient to establish whether a method is applicable in the context of encryption.

Regarding cryptographic security, Shannon's work [5] shows that the highest level of security is reached when applying a secure cipher to a redundancy free plain text. Current image/video codecs exploit redundancy for compression and thus we can consider a bit stream to be an almost redundancy free plain text in the sense of Shannon. Consequently, for maximal security, the encryption of the entire bit stream with a state-of-the-art cipher, such as AES, would suffice ("conventional encryption"). Lookabaugh and Sicker [6] showed that selective encryption is sound and demonstrated its relation to Shannon's work. However, [7] showed that side information can compromise security.

However, there are application scenarios which make it necessary to move away from full encryption. Methods which do not utilize full encryption of the underlying data are called Lightweight/Soft/Partial/Selective Encryption. Specifically, selective encryption is the application of an assumed secure encryption method to a *selected* part of the plain text. In selective encryption the *encryption* part is assumed to be secure, e.g., by using AES. The final security of the selective encryption comes then from the *selection* part. What is evaluated in order to gauge the final security of the selective encryption is, to what extent the information left in plain text can be used to reconstruct an image or video.

Furthermore, an attack on the selective encryption method does not come from attacking the *encryption*, but from attacking the *selection*. This is usually done by using knowledge about the original format of the image/video. An attack is usually based on removing the negative impact on quality by the, essentially random, signal introduced by the encryption. This is typically done by replacing the random signal by a signal which introduces the least amount of error into the final decoding. In order to do so, very specific knowledge about the containing format has to be exploited, and there is usually only a single method to go about this, i.e., *the attack*.

* Corresponding author.
E-mail addresses: hhofbaue@cosy.sbg.ac.at (H. Hofbauer), uhl@cosy.sbg.ac.at (A. Uhl).

http://dx.doi.org/10.1016/j.image.2016.05.001

Another, (implicit) assumption about the selective encryption method under test is the *format compliance*. Format compliance requires an encrypted bitstream to be decodable by a standards compliant decoder. In other words, format errors may not be introduced by the encryption. Thus, in the following, when we refer to an encrypted image/video we mean the image or video that results from decoding an unattacked encrypted bitstream.

The notion of security in selective encryption is different from the traditional notion of security: First, we knowingly leave information in plain text to retain format compliance; second, the focus is on content security not information security, i.e., the content should be secure (to some defined extent), while information about the content might be allowed to leak. In order to be able to discuss the exact notion of security in such non-conventional encryption schemes, we need to distinguish distinct application scenarios of encryption schemes for visual data:

Confidentiality encryption: Means MP security (message privacy). The formal notion is that if a system is MP-secure an attacker cannot efficiently compute any property of the plain text from the cipher text [8]. This can only be achieved by the conventional encryption approach.

Content confidentiality: This is a relaxation of confidential encryption. Side channel information may be reconstructed or left in plaintext, e.g., header information, packet length, but the visual content must be secure in the sense that it must not be intelligible/discernible [9].

Sufficient encryption: Means we do not require full security, just enough security to prevent abuse of the data. The content must not be consumable due to high distortion (e.g., for DRM systems) by destroying visual quality to a degree which prevents a pleasant viewing experience or destroys the commercial value. This implicitly refers to message quality security (MQ), which requires that an adversary cannot reconstruct a higher quality version of the encrypted material than specified for the application scenario [10].

Perceptual/transparent encryption: Means we want consumers to be able to view a preview version of the video but in a lower quality while preventing them from seeing a full version. As an example: this can be used in a pay per view scheme where a lower quality preview version is available from the outset to attract the viewers interest, q.v., [11]. The difference between sufficient and transparent is the fact that there is no minimum quality requirement for sufficient encryption. Encryption schemes which can do sufficient encryption cannot necessarily ensure a certain quality and are thus unable to provide transparent encryption.

Given these different application scenarios it is clear that depending on the goal, a security metric has to fulfill different roles. For example, under the assumption of sufficient encryption, a given security metric would have to evaluate which quality is low enough to prevent a pleasant viewing experience. In contrast, for the transparent encryption case, a metric not only has to assess whether the quality of an image or video is low enough, but also whether the quality is high enough to be useful to attract interest. When it comes to content confidentiality the question of quality is no longer applicable. Content confidentiality requires that image content must not be identified by human or automated recognition. This requirement also has to be maintained for any part of the image. Image metrics, in general, do not deal with such questions but rate the overall image quality, the question of intelligibility is usually not covered at all. A drastic example would be an image where only a small part of the image is partly visible. Classical metrics would judge the whole image and consequently would attribute a high security, even though a part of the image is still recognizable which contradicts content confidentiality. Still, it has to be pointed out that content confidentiality can have different forms. To prevent a face recognition scheme from working properly it is sufficient to protect any facial information in a surveillance video, while humans could still be identified in such a video by using gait recognition. Furthermore, if the appearance of a person has to be concealed entirely, a much stronger extent of protection (i.e., higher security) is required. Finally, confidential encryption cannot be solely assessed with security metrics since the scope goes beyond assessing security based on the visual appearance only. Furthermore, we should note that the application of security metrics on video is performed at a frame by frame basis in the literature. We will adopt this model but should note that for the discussion of confidential encryption motion data is of importance, e.g., in [12] it was shown that a replacement attack combined with motion information can reveal the content of a scene even though the visual content of every frame is encrypted.

Consequently, depending on a given application scenario different properties are required from a security metric and different approaches to construct such a metric might perform better or worse for some applications scenarios. This dependence on the evaluation goal of a security metric is hardly ever discussed in the papers introducing a metric. Sufficient and transparent encryption scenarios have a clear and distinct link to the traditional notion of (low) visual quality, while it is highly questionable or at least doubtful if content confidentiality can be assessed by the classical quality notion. While the lack of relation to spatial areas of most security metrics could be compensated in the design to provide locally varying results, the lack of relation to intelligibility in general can probably not be easily resolved.

For both, security metrics and regular image metrics, in the literature we do not find any evaluation whether a given metric can perform the claimed function or how such an evaluation correlates to actual security. However, for regular image metrics it is well known that the correlation with human observations over the full range of possible quality (from high to low quality) does not imply a good performance on a given subset. More specifically, it was pointed out recently that most image metrics perform very poorly for the low quality range ([13]–using the low quality end of the LIVE database). For security metrics, not even this question has been covered so far.

In this paper, we will try to remedy this situation by giving an overview of requirements regarding security goals and formulating these requirements into a testing methodology. Based on this methodology we will evaluate the various security metrics in the literature as well as applicable conventional image metrics. However, we will not deal with every application scenario equally explicitly. We will only make a first step to cover the content confidentiality scenario. The main reason for this is a lack of ground truth. It is not obvious how to generate ground truth for this scenario since there is a disparity between how an image metric works and what is necessary to evaluate content confidentiality. Image metrics, and as an extension security metrics, measure the quality of an image respective to human judgement. This works well for high quality images but suffers for low quality images where human observers can have difficulties differentiating between the severity of an impairment. Thus the methodology to systematically generate ground truth based on human observation needs to be changed for content confidentiality which is not in the scope of this paper. On the other hand, for the image quality-related scenarios (sufficient and transparent encryption), ground truth data is available, in the form of image impairment databases with mean opinion scores (MOS) based on a number of human observations.

In the following we will motivate and introduce a methodology

which allows to measure desirable traits of a security metric. This methodology can, and should, be used to asses newly developed metrics.

To motivate this, we will give an overview of security metrics used in the literature in Section 2. We will not analyse why metrics fail and the presented counterexamples should be seen as a proof by contradiction rather than a exhaustive list of failings. Note that this is a question of generality, we do not dispute that certain metrics are a good fit for specific encryption types.

In Section 3 we will describe what is expected of security metrics and the methodology how security metrics can be evaluated. This is done by providing certain desirable traits of a security metric, a motivation for the traits and a formal description of the traits to allow measuring them.

In Section 4 we will present the evaluation of state-of-the-art metrics and measure their ability to perform as security metrics. This is done based on the traits introduced in Section 3 and shows the usefulness of these traits to find shortcomings of current metrics and to prevent the same shortcomings in future metrics.

Section 5 will conclude the paper and give a list of state-of-the-art reference metrics and their usability as security metrics.

2. Overview of security metrics

In this section a brief overview of security metrics will be given. The metrics discussed are taken from the recent literature and are specifically designed to ascertain whether the image quality after encryption is sufficiently reduced. The metrics given in this section are discussed as general security metrics, i.e., not limited to the specific method with which they were designed together. References to the original work will be given for each security metric as well as some examples where the metric fails to assess given example images as would be expected from a general security metric. The SSIM and PSNR are also included in this overview. Even though they were not designed to be security metrics, they are frequently used as such, e.g., [3,14–16] (SSIM), [4,17–19] (PSNR) and [1,2] (SSIM and PSNR).

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 judgement is somewhat lacking even for high and medium quality [20]. Fig. 1 illustrates the performance of the PSNR metric on samples from the IVC-SelectEncrypt [21] database (see Section 3).

Structural Similarity Index Measure (*SSIM*): The structural similarity index measure (SSIM) [22] 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 $[0, 1]$. Fig. 2 illustrates the performance of the SSIM metric on samples from the IVC-SelectEncrypt database.

Edge Similarity Score (*ESS*): The edge similarity score (ESS) was introduced in [23] and uses localized edge direction information to compare two images. The ESS is a quality metric, a high metric score reflects a high quality, which gives a score in the range $[0: 1]$. Fig. 3 illustrates the performance of the ESS metric on the foreman sequence when encryption according to [24] is applied in comparison to white noise.

Luminance Similarity Score (*LSS*): The luminance similarity score (LSS) was introduced in [23] and uses localized luminosity information to compare two images. The LSS is a quality metric, a high metric score reflects a high quality, which gives a score in the range $[-8.5: 1]$. Fig. 4 illustrates the performance of the LSS metric on the foreman sequence when encryption according to [25] is applied in comparison to noise.

Neighborhood Similarity Degree (*NSD*): The neighborhood similarity degree metric (NSD), introduced in [26], uses local pixel similarity correlation between original and impaired image. The NSD depends on two parameters, one to define the region for pixel similarity correlation (d) and one to define the similarity threshold (m). The parameters m and d were set to the same values as in the experiments in [26], i.e., $m=5$, $d=3$, border extension is done by repeating the last border pixel. The NSD is an impairment metric, a high metric score reflects a low quality, which gives a score in the range $[0: 1]$. Fig. 5 illustrates the performance of the NSD metric on samples from the IVC-SelectEncrypt database.

Local Entropy (*LE*): The local entropy metric was introduced in [27] (LE), it is a no reference metric operating only on an impaired image. The LE metric uses the average of normalized localized entropy scores, on 8×8 blocks, as image quality predictor. The LE is an impairment metric, a high metric score reflects a low quality, which gives a score in the range $[0: 1]$. Fig. 6 illustrates the performance of the LE metric on samples from the IVC-SelectEncrypt database.

Local Feature Based Visual Security (*LFBVS*): The local feature based visual security metric (LFBVS) was introduced in [28] and utilizes localized edge and luminance features which are combined and weighted according to error magnitude, i.e., error pooling. The LFBVS is an impairment metric, a high metric score reflects a low quality, which gives a score in the range $[0: 1]$. Fig. 7 illustrates the performance of the LFBVS metric on the silent sequence when encryption according to [24] is applied in comparison to white noise.

All the given image metrics, with the exception of the local entropy metric (LE), are full reference metrics, meaning they utilize information from the original and comparison (encrypted) image to calculate an assessment of the visual similarity. The local

Fig. 1. PSNR metric scores for images from the IVC-SelectEncrypt database. According to PSNR images (a) and (b) are of the same quality and (c) is of much higher quality, i.e., less secure than (a) and (b).

Fig. 2. SSIM metric scores for images from the IVC-SelectEncrypt database. According to SSIM images (a) and (b) are of the same quality and (c) is of much lower quality, i.e., more secure than (a) and (b).

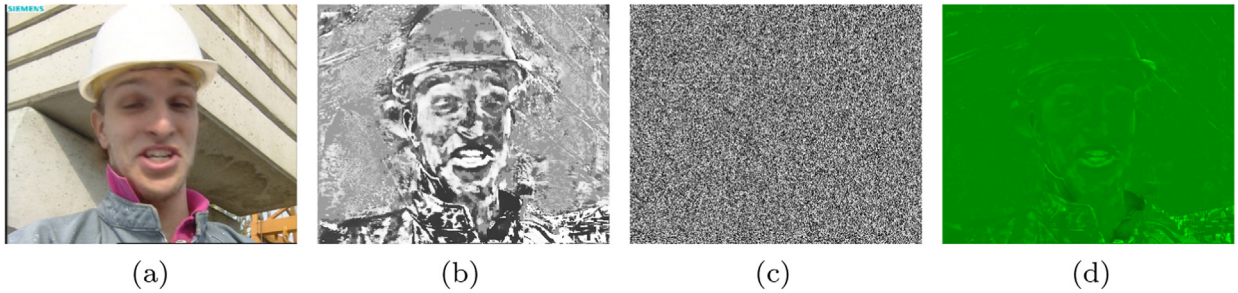


Fig. 3. ESS comparison for frame 80 of the foreman sequence (a). ESS judges the white noise (c) to be of higher quality than the residual information from the encrypted frame (d). In order to show the amount of information actually retained in the encrypted frame a post processed version is also shown (b).

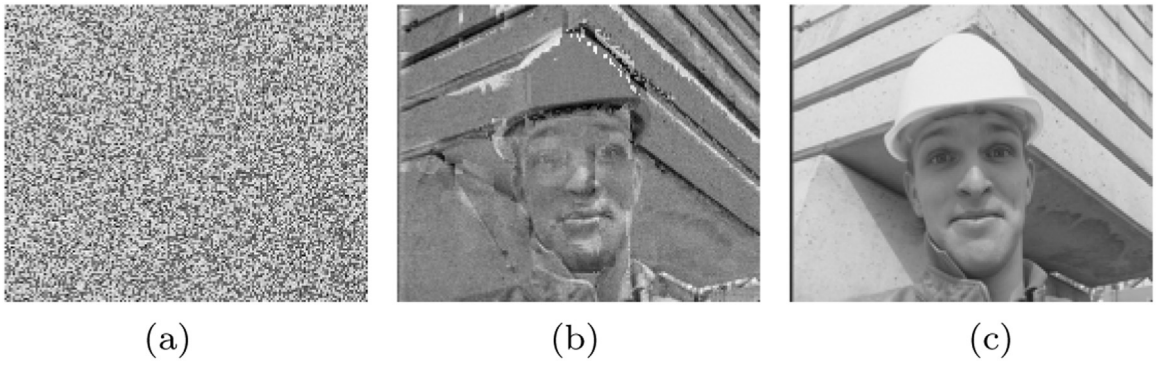


Fig. 4. LSS comparison for a frame of the foreman sequence (c). LSS judges the noise (a) to be of higher quality than the residual information from the encrypted frame (b).

entropy metric by Sun et al. is a no reference metric, i.e., it utilizes only the impaired image to judge the resulting quality. By measuring entropy, LE can also be interpreted to assess the encrypted image compared to random noise (which exhibits maximal LE). Since all of the given security metrics are proposed to be general we will not differentiate between full- and non-reference metrics in the following but compare them solely on the task they are supposed to solve.

From the description of the various security metrics it can be seen that a wide range of approaches exist, from metrics targeting signal properties, e.g., PSNR which targets noise, to LE which targets local entropy, metrics which use higher level information, e.g., NSD or ESS which use a form of object detection (mostly based on edges), to metrics which use information about the HVS to improve their performance, e.g., SSIM or LFBS which use simulation of the fovea centralis and error pooling respectively. However, for

Fig. 5. NSD metric scores for images from the IVC-SelectEncrypt database. According to NSD images (a) and (b) are of the same quality and (c) is of much lower quality, i.e., more secure than (a) and (b).

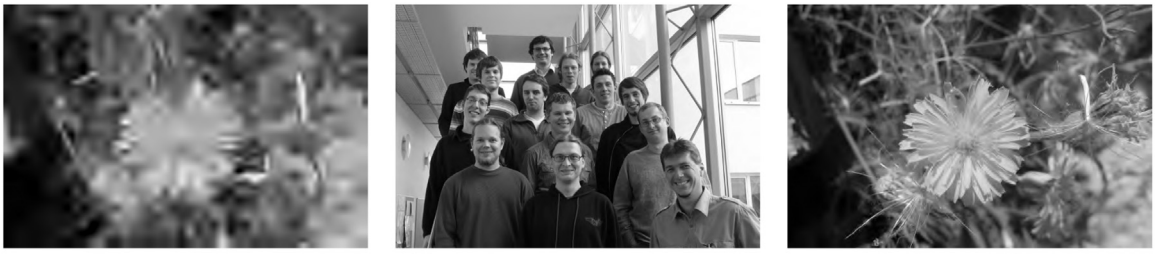


Fig. 6. LE metric scores for images from the IVC-SelectEncrypt database. According to LE images (a) and (b) are of the same quality and (c) is of much lower quality, i.e., more secure than (a) and (b).

every security metric the accompanying figure demonstrates an obvious fault in the performance of the given metric. Such examples can be found for every metric of course, the question we will try to answer in the following sections is whether the demonstrated fault is singular or systemic.

In order to gauge the effectiveness of a dedicated security metric it is useful to compare them to regular metrics. To facilitate a fair comparison, three recent metrics, in addition to PSNR and SSIM, are chosen and included in the evaluation. The local edge gradient image metric (LEG) [29] shows a good correlation with human judgement and is reasonably fast to compute such that it can be used even under time constraints. The visual information fidelity (VIF) [30] and CPA1 [31] outperform the SSIM and LEG in regard to correlation with human judgement but are a lot slower to compute [29]. A second reason to include those metrics is to gauge whether they can be used as security metrics. The LEG and VIF are quality metrics with a score in the range $[0, 1]$, and the CPA1 is an impairment metric with a score in $[0, \infty]$.

3. Evaluation methodology

In this section we outline (1) our evaluation methodology, (2) the reason to use this methodology, and (3) the application scenario(s) which can be assessed by employing a certain methodology. A discussion of the desired outcome from these tests for security metrics is also provided. This section is the guideline of how security metrics and image metrics are evaluated for the use as security metrics in Section 4.

3.1. Application domain

Selective encryption methods, in a majority of cases, aim at format-compliant encryption, i.e., encryption in such a way that the media can be decoded by a standards compliant decoder without error. Frequently, security metrics are applied on a direct decoding of the encrypted bitstream. This can have adverse effects since encryption introduces noise which can hide plain text data and consequently a security metric might judge that an encryption method is more secure then it actually is. There are a number of options how a security metric can be applied, as illustrated in Fig. 8. All security metrics under evaluation are applied on a direct decoding of the encrypted bitstream (which we will denote as the *encrypted domain*) by the authors in the corresponding original papers. Other options would be to attack the encryption method in a way which does not break it but reduces the obfuscation of the plain text data in the encrypted domain (denoted *extracted* in the figure). Such attacks usually utilize knowledge about the bitstream rather than the encryption method (other than location). Typical attacks would be error concealment and replacement attacks against selective encryption schemes [32]. In general, such attacks replace the encrypted content with bitstream elements that are statistically known to introduce less noise than the random elements produced by encryption. Another possibility would be to utilize post processing to further help the metric detect residual information (denoted *processed* in the figure).

In the evaluation we only handle the difference between security metrics applied directly in the encrypted domain versus application in the extracted domain. The post processing step is only provided to better highlight the remaining information in an image. It cannot be directly used as an application domain since the post processing step is either specific to the encryption, in which case it should be included in the attack, or specific to the metric, in which case it should be included in the security metric. Post processing, in general, can influence different metrics in various ways, an example of this is given in Fig. 9 where post processing increases ESS and, at the same time, decreases the SSIM. This means that it is never actually possible to determine an optimal postprocessing method which would be required in a sensible assessment.

In order to test whether a security metric can operate in the encrypted domain a number of well known video sequences have been encrypted for three target qualities, utilizing EZBC encryption methods described in [24]. To generate the different target qualities the selective encryption was applied to either all I-frames (low quality), low frequency bands of all frames (medium quality), and

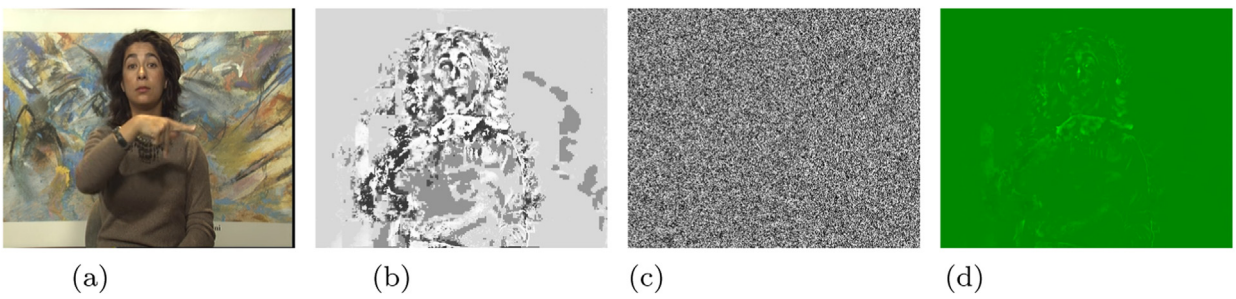


Fig. 7. LFBVS comparison for frame 80 of the silent sequence (a). LFBVS judges the residual information from the encrypted frame (d) to be about the same as the information contained in white noise (c). In order to show the amount of information actually retained in the encrypted frame a post processed version is also shown (b).

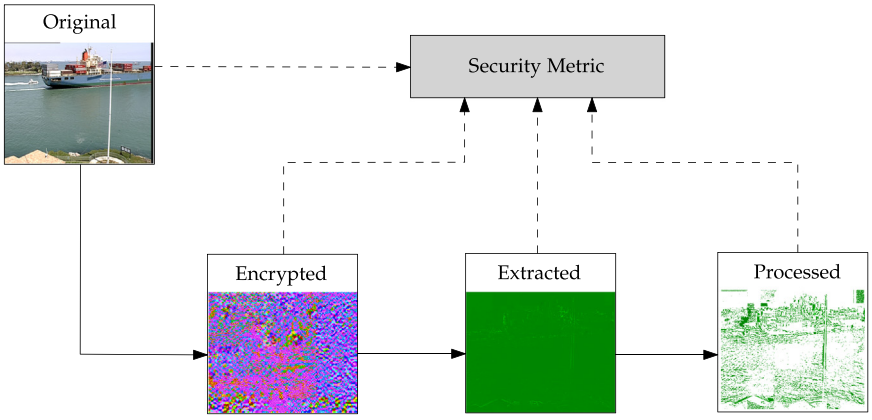


Fig. 8. The possible domains which can be used by a metric to compare to an original image. Either the direct output of an encrypted bit stream for format compliant encryption or an extraction of the plain text data which minimizes the disruptive effect of the encrypted data on the resulting bit stream. Another possibility would be post processing to further accent the residual plain text data contained in an image.

high frequency bands of all frames (high quality). Fig. 10 illustrates the quality targets of the encryption process. The targets where chosen to contain high, medium and low residual information. Under the assumption that a security metric can operate in the encrypted domain it should be able to reliably order the encrypted frames of each sequence for every sequence from highest to lowest quality.

We generate two sets of ground truth, one for the high-medium and one for the medium-low quality comparison, containing ordered (based on quality) tuples for each frame. For a sequence S containing frames $s^1, \ldots, s^{N_S}$ and high (S_H), medium (S_M) and low (S_L) quality versions we generate two sets of ordered (by quality) tuples $D_{HM}(S) = \{(s_H^1, s_M^1), \ldots, (s_H^{N_S}, s_M^{N_S})\}$, and likewise for D_{ML}.

Then a tested metric should also generate two sets with ordered tuples based on the estimated quality provided by the metric
$M(o, c) \mapsto \mathbb{R}$: $Q_{HM}(S, M) = \{q(M, s^1, s_H^1, s_M^1), \ldots, q(M, s^{N_S}, s_H^{N_S}, s_M^{N_S})\}$,
where

$$q(M, o, c_1, c_2) = \begin{cases} (c_1, c_2) & \text{if } M(o, c_1) \geq M(o, c_2), \\ (c_2, c_1) & \text{otherwise,} \end{cases}$$

and likewise for $Q_{ML}(S, M)$.

Based on these four sets the ability of a metric to order in a given domain can be estimated by

$$O_M(S) = \frac{1}{2}\frac{|Q_{HM}(S, M) \cap D_{HM}(S)|}{|D_{HM}(S)|} + \frac{1}{2}\frac{|Q_{ML}(S, M) \cap D_{ML}(S)|}{|D_{ML}(S)|}. \quad (1)$$

values of 1 (100% correct ordering) or 0 (0% correct orderings, i.e., 100% correct reverse ordering) show a high ability of the metric to order correctly in the given domain. The difference between 1 and 0 is that image metrics can either measure similarity of images, i.e., quality metrics, resulting in a normal ordering or the difference between images, i.e., impairment metrics, resulting in a reverse ordering. Results around 0.5 are akin to random decisions and reflect the inability of the tested metric to perform in this domain.

Furthermore, since the low quality range chosen is in (or at least close to) the domain of content confidentiality, i.e., the rightmost column in Fig. 10 does not exhibit intelligible visual content, this setting also serves as an indication whether an image metric might be useful for content confidentiality. While a good performance on this evaluation does not necessarily mean a image metric is qualified for content confidentiality, a low performance is a strong indicator that the metric is unfit for this task. Based on the information which parts of the data have been encrypted and the entirely evident differences in visual appearance, ground truth is out of question here.

Note also that this is a proof by contradiction. That is, if a metric

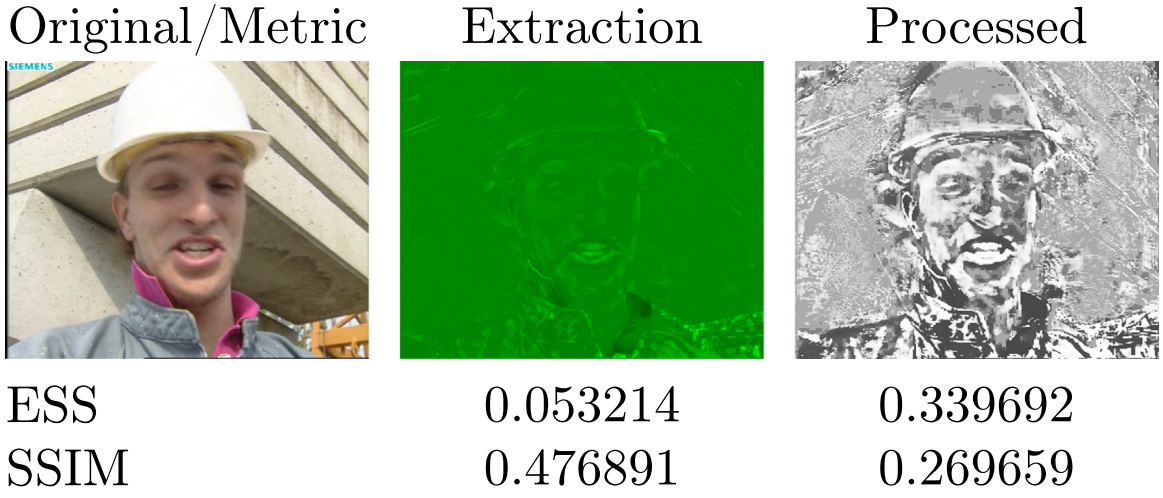


Fig. 9. An example of how post processing influences different metrics.

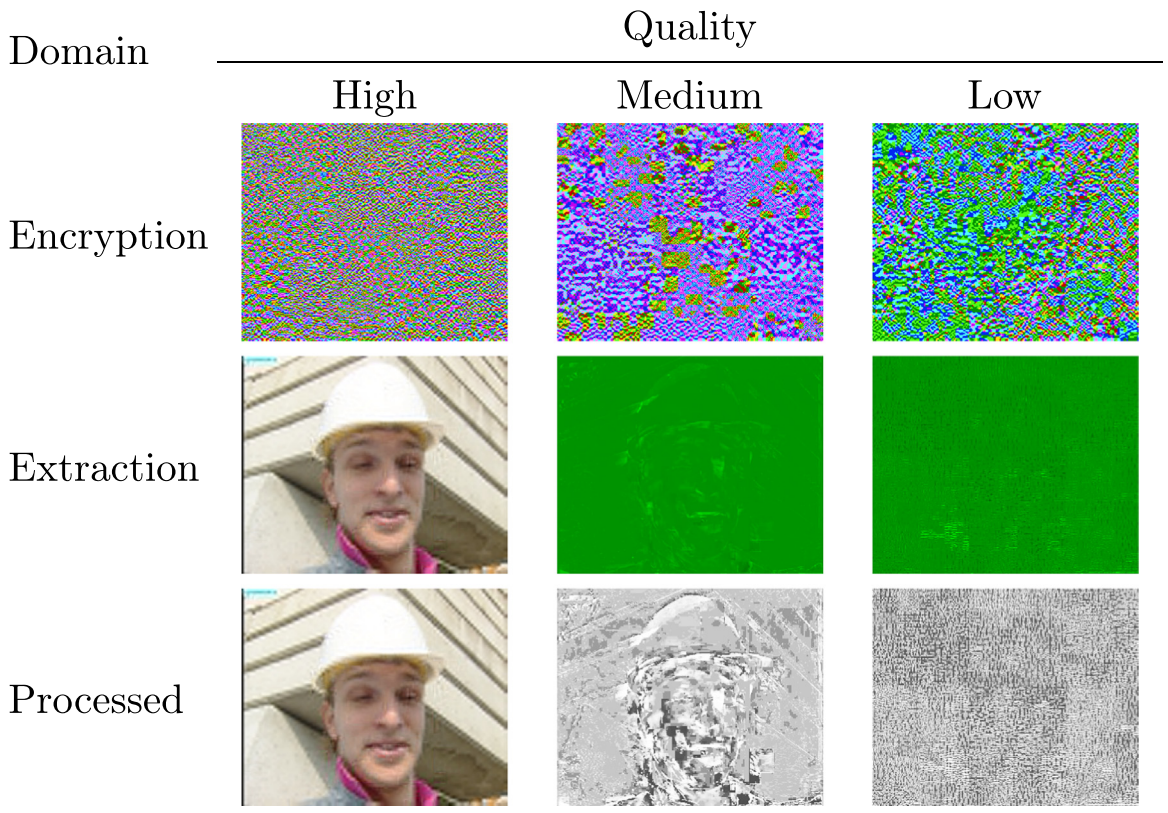


Fig. 10. A sample of the quality ordering test set based on the foreman sequence. Samples from the high, medium and low quality sequences are shown in the encryption, extraction and processed domain.

claims to be able to generally work in the encrypted domain and we can demonstrate that there are cases where it does not, then this claim has been falsified.

3.2. Correspondence to HVS: confidence

Besides the encryption application scenarios where a certain quality is required (sufficient and transparent encryption), further examples for the importance of the quality notion are watermarking where the resulting quality should not be below a certain threshold, and of course, lossy compression. However, the notion of quality in this cases is not as straightforward as it seems. On the one hand we use the term quality in the context of the human visual system (HVS), i.e., how a person consuming the content would judge the quality. On the other hand, the term quality can refer to the score returned by a (security) metric which is tied to the quality in the HVS sense. This relation is not exact and it is not inherently clear how to choose a metric which correlates to the HVS quality which is targeted, although in practice Algorithm 1 is usually applied.

Algorithm 1. Method for finding a target metric score based on a target HVS quality.

1: Chose a source image.
2: Alter the image until it fits the perceived target quality.
3: Apply the security metric on the altered and original image, the resulting metric score is the target quality.

While this results in a target quality which can be used, we know nothing about how well this score actually reflects the human judgement, since it is well known that the correlation between human judgement and image metrics is not perfect. In other words, how confident can we be in the choice of image metric score in relation to the perceived quality?

In order to evaluate this, well known databases which contain impaired and encrypted images and the perceived quality, in the form of mean opinion scores (MOS), will be used. The databases contain a set of points p representing impaired images with associated values p_v for metric value and p_d for MOS value, ordered from lowest to highest quality. Based on a target MOS quality score D two values can be calculated, Fig. 11 illustrates this.

Zero false negative: $V_{min}(D)$ refers to the metric value for which the following holds: $p_d > D \Rightarrow p_v > V_{min}(D)$. That is if the metric score is below $V_{min}(D)$ we are sure that the perceived quality is below the MOS quality score (D).

Zero false positives: $V_{max}(D)$ refers to the metric value for which the following holds: $p_v > V_{max}(D) \Rightarrow p_d > D$. That is if the metric score is above $V_{max}(D)$ we are sure that the perceived quality is

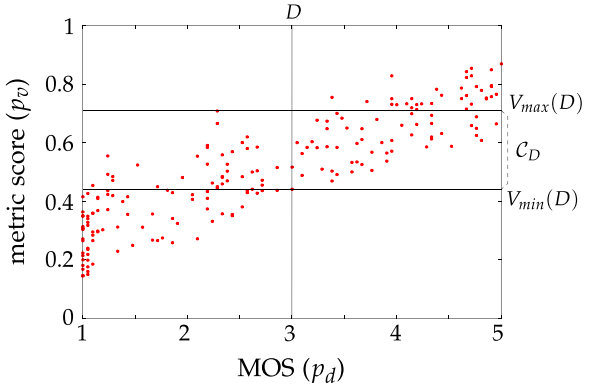


Fig. 11. Illustrations of zero false negative $V_{min}(D)$, zero false positives $V_{max}(D)$ and confidence C_D for a MOS value of $D=3$ is shown based on the IVC-SelectEncrypt database and the LEG metric.

above the MOS quality score (D).

This also means that if a target metric quality score ${p_v}^t$ is obtained as given by Algorithm 1 we are assured that $V_{min}(D) \geq p_v^t \geq V_{max}(D)$.

Thus we can define the confidence C_D for a metric score based on a given perceived quality D as $C_D := |V_{max}(D) - V_{min}(D)|$. A confidence score over the full perceived quality range can be given as

$$C = \frac{1}{\#S} \sum_{D \in S} C_D,$$

where S is the set of distinct MOS samples from the database.

Also note that we can interpret C as the average over C_D, $\mu_{D \in S}(C_D)$, and consequently we can also calculate the standard deviation, $\sigma_{D \in S}(C_D)$. The reason for calculating σ is to estimate how stable the confidence range is over the whole range of visual quality. This has to be taken into account since it is well known that image metrics exhibit different correlation to human judgement depending on the quality range, e.g., [13].

For security metrics, and image metrics in general, the lower $\mu(C_D)$ and $\sigma(C_D)$ the better Algorithm 1 can be used to estimate a target image quality metric score.

Furthermore, since the signal is reduced to statistical components it is also of interest which shape the signal takes in conjunction with $\mu(C_D)$ and $\sigma(C_D)$. The shape, together with the monotonicity (see Section 3.3), can be used to indicate a possible application scenario for a security metric, essentially whether the security metric can be used for all quality ranges or only on high/low quality applications.

By shape of the signal we mean the distribution of outliers, where we define outlier based on the z score[1] of a data point D as

$$z_D = \frac{C_D - \mu(C_D)}{\sigma(C_D)}.$$

we will define *high outliers* as outliers with $z_D < -1$ and likewise *low outliers* as outliers with $z_D > 1$, indicating a higher and lower confidence respectively. Based on the distribution of high and low outliers we can specify the shape of the signal as follows.

- A signal is *stable* if there are no outliers. That is, if $-1 \geq z_D \geq 1$ holds for all $D \in S$.
- A signal is *biased* if it consist of two clearly separable parts, where one exhibits a good confidence score (C_D small) and the other a bad confidence score (C_D large). We say the signal is biased towards the quality where the good confidence score is located, i.e., where the metric is performing well.
  That is, A signal is biased if there exists a D_t such that $z_D < -1 \Rightarrow D < D_t$ and $z_D > 1 \Rightarrow D > D_t$ or $z_D < -1 \Rightarrow D > D_t$ and $z_D > 1 \Rightarrow D < D_t$. If a low D indicates a high quality we specify the shape to be *biased towards high quality* if $z_D < -1 \Rightarrow D < D_t$ or *biased towards low quality* if $z_D > 1 \Rightarrow D < D_t$. If a low D indicates low quality the definition is switched accordingly.
- A signal which is neither stable nor biased is considered *unstable*.

3.3. Correspondence to HVS: monotonicity

What is required from image metrics in general is monotonicity with regard to human observations. That is, if an image metric decides that image A is of better quality than image B a human observer should also prefer image A over image B. This is akin to correlation but since the human visual system is not a linear system regular linear correlation is meaningless. Thus in order to ascertain the correlation of an image metric and human observations the notion of monotonicity is utilized. Rank order correlation, which essentially judges the monotonicity of the signals, is most often used, usually in the form of Spearman's rank order coefficient (SROC) [33] or Kendall Tau (τ) [34].

Hofbauer and Uhl [13] pointed out that the correlation of an image metric over the full quality range does not imply that a high correlation is achieved for the low quality range. This is especially important for security metrics since certain application scenarios specifically target the low quality range of images, e.g., sufficient encryption. We cannot confine the evaluation to the low quality range since there are also applications for higher quality, i.e., transparent encryption. Also note that this is a dual property to the confidence in the sense that for the confidence we evaluate the relation of choosing a MOS value and evaluating the range of metric scores which can potentially fall onto this MOS value. Monotonicity is evaluated on specific sets of impairment and looks at how well an increase in metrics score reflects an increase in the MOS.

To properly evaluate security metrics for all encryption scenarios we will evaluate them using a high quality, a low quality and a full quality range dataset. The reason to also include the high quality range is to be inclusive in terms of possible application scenarios. For example the upgrade to high definition quality from PAL/NTSC quality is just as valid in terms of application scenarios as from hand held quality to PAL/NTSC quality. As basis for the evaluation we will use well known databases which contain mean opinion score of human judgement over different impairments. The SROC will be used for evaluation purpose, as is current best practice for metric evaluation.

From security metrics we would expect a high correlation with human judgement for the low quality range. While the low quality range is often the target of encryption some transparent encryption schemes could target a higher quality, consequently, a good correlation with human judgement on the high quality range is also desirable.

4. Evaluation

In this section we present the results of the evaluation process as detailed in Section 3. Each evaluation contains a short description of the test data, the actual data from the evaluation and a discussion. We remark that the discussion is focussed on the evaluation methodology rather than the details of why and how a metric fails a test. Discussing the latter would require an in-depth description of how the metrics work, together with the explicit details about the test set which is out of scope of this paper.

With respect to implementations, we used our own code[2] for LSS, ESS, LE, NSD, LFBVS, SSIM and PSNR. For VIF, we used the implementation from the "MeTriX MuX Visual Quality Assessment Package",[3] version 1.1. For CPA1, we used the MATLAB implementation provided by Florent Autrusseau.[4]

4.1. Evaluation of the application domain

In order to evaluate the extraction versus encrypted domain applicability of metrics, we used a number of standard sequences: akiyo, bus, coastguard, container, flower, foreman, mobile, news, silent, tempete and waterfall.[5] The ordering, as discussed in

[1] NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, October 2013.

[2] http://www.wavelab.at/sources/VQI

[3] http://foulard.ece.cornell.edu/gaubatz/metrix_mux/

[4] http://www.irccyn.ec-nantes.fr/~autrusse/Softwares.html

[5] Available for example at http://media.xiph.org/video/derf

Section 3.1 above is performed on a frame by frame basis and averaged over all the frames in a given sequence. Additionally we provide the average over all sequences in order to simplify the comparison. Table 1 shows the results for the encryption and extraction domain. The optimal result would be for a metric to perform equally well independent of the application domain.

From the overall averages it can be clearly seen that the performance in the encrypted domain is worse than in the extraction domain with the exception of LE. While LE performs better in the encrypted domain than in the extraction domain, the performance is still very low. In the extraction domain, most metrics still perform poorly; only LEG, SSIM, VIF, CPA1 and, to a lesser degree, ESS exhibit good performance. Aside from the latter, all other security metrics exhibit extremely poor performance.

This allows to conclude that *application in the encrypted domain should not be performed* (*although*, *it is routinely done in all corresponding papers*).

When looking closer at the detail information from the extraction domain, some interesting effects can be observed. While the performance of LEG, SSIM and VIF are consistently good, we notice that there are cases where even a metric that shows good performance can have problems. In the actual case the performance of the ESS is significantly reduced for the waterfall sequence and the CPA1 shows poor performance on the bus sequence. It stands to reason that there are other, untested, sequences which would lead to similar reduced performance for the LEG, SSIM and VIF.

4.2. Evaluation of confidence

In order to evaluate the confidence, databases with either pertinent content, i.e., encrypted images, or a large dataset with distortions are optimal. The IVC-SelectEncrypt database [21] contains various instances of JPEG 2000 transparent encryption, using different encryption techniques, and is a useful tool for evaluating confidence with respect to encrypted images. It is the only database available containing encrypted visual data and corresponding MOS. The test sets contained in the IVC-SelectEncrypt database (and their abbreviation) are traditional encryption (trad), truncation of the code stream (trunc), window encryption without error concealment (iwind_nec), window encryption with error concealment (iwind_ec), and wavelet packet encryption (res), for detailed information see [35]. However, the IVC-SelectEncrypt database has a rather small set of impairments, i.e., five per test set, and is focused on JPEG 2000 only.

In order to get a more diverse view on the confidence of metrics we additionally utilize the LIVE database [36] to supplement the IVC-SelectEncrypt database of encrypted images. While the LIVE database does not contain encrypted images, the quality range of the images reaches from high to low quality; this makes it at least relevant for transparent encryption where a certain target quality is required. Furthermore, images in the low quality range of the LIVE database exhibit strong distortions which can be equated to encrypted images in the sense that strong distortions mask a lot of the visual information. Consequently, the distorted images can be used to assess how well a metric can identify information contained in a distorted/encrypted image; this is exactly the property that we want from security metrics. An example of these strong distortions is shown in Fig. 12 which contains an encrypted image from the IVC-SelectEncrypt database as well as heavily distorted versions of images from the LIVE database. These examples illustrate that the LIVE database contains not only images which are similar to the IVC-SelectEncrypt database in terms of content masking, but also images which are clearly in the quality realm of sufficient encryption. The test sets contained in the LIVE database (and their abbreviation in plots and figures) are JPEG 2000 compression (jp2k), JPEG compression (jpeg), white noise (wn), Gaussian blur (gblur), and bit errors in JPEG2000 bit stream transmission over a simulated fast fading Rayleigh Channel (fastfading), for detailed information see [37].

What we tried to do is to break it down the complex relation of the confidentiality as far as possible, to facilitate comparison,

Table 1
Results for the ordering given by Eq. (1) (Section 3.1) of high versus medium and medium versus low quality sequences in the encryption and extraction domain. The averages over all the frames in a sequence as well as the average over all sequences is shown per image metric.

	LEG	SSIM	LSS	ESS	PSNR	LFBVS	LE	NSD	VIF	CPA1
Results for the encryption domain (Q_M)										
akiyo	0.168	0.500	0.496	0.441	0.527	0.500	0.609	0.481	0.449	0.496
bus	0.398	0.383	0.402	0.332	0.555	0.516	0.895	0.426	0.383	0.250
coastguard	0.500	0.391	0.340	0.434	0.582	0.590	0.754	0.422	0.508	0.254
container	0.246	0.484	0.207	0.391	0.348	0.500	0.777	0.563	0.586	0.035
flower	0.191	0.430	0.527	0.410	0.445	0.488	0.996	0.398	0.488	0.395
foreman	0.289	0.481	0.492	0.410	0.508	0.504	0.547	0.453	0.500	0.481
mobile	0.340	0.481	0.500	0.305	0.457	0.512	0.688	0.492	0.453	0.473
news	0.430	0.500	0.395	0.418	0.742	0.500	0.500	0.457	0.363	0.496
silent	0.270	0.332	0.242	0.102	0.250	0.602	0.992	0.375	0.320	0.148
tempete	0.414	0.481	0.500	0.395	0.500	0.500	0.742	0.481	0.449	0.465
waterfall	0.422	0.492	0.488	0.461	0.715	0.508	0.539	0.512	0.492	0.496
average	0.334	0.450	0.417	0.373	0.512	0.520	0.731	0.460	0.454	0.363
Results for the extraction domain (Q_M)										
akiyo	1.000	1.000	0.516	0.941	1.000	0.500	0.500	0.500	0.945	0.094
bus	0.996	1.000	0.512	0.969	0.992	0.500	0.500	0.461	0.981	0.234
coastguard	1.000	1.000	0.508	0.992	0.996	0.457	0.500	0.246	1.000	0.008
container	1.000	1.000	1.000	0.941	0.000	0.500	0.500	0.500	1.000	0.004
flower	1.000	1.000	0.953	0.926	0.063	0.418	0.500	0.473	0.984	0.094
foreman	1.000	1.000	0.531	0.996	0.500	0.500	0.500	0.500	0.996	0.000
mobile	0.984	1.000	0.922	0.977	0.434	0.500	0.500	0.410	0.984	0.020
news	1.000	1.000	0.500	0.996	1.000	0.500	0.500	0.500	0.996	0.000
silent	1.000	1.000	0.902	0.981	0.578	0.500	0.500	0.500	1.000	0.000
tempete	1.000	1.000	0.570	0.992	1.000	0.500	0.500	0.481	1.000	0.000
waterfall	0.957	0.992	0.695	0.828	0.930	0.500	0.500	0.246	0.981	0.008
average	0.994	0.999	0.692	0.958	0.681	0.489	0.500	0.438	0.988	0.042

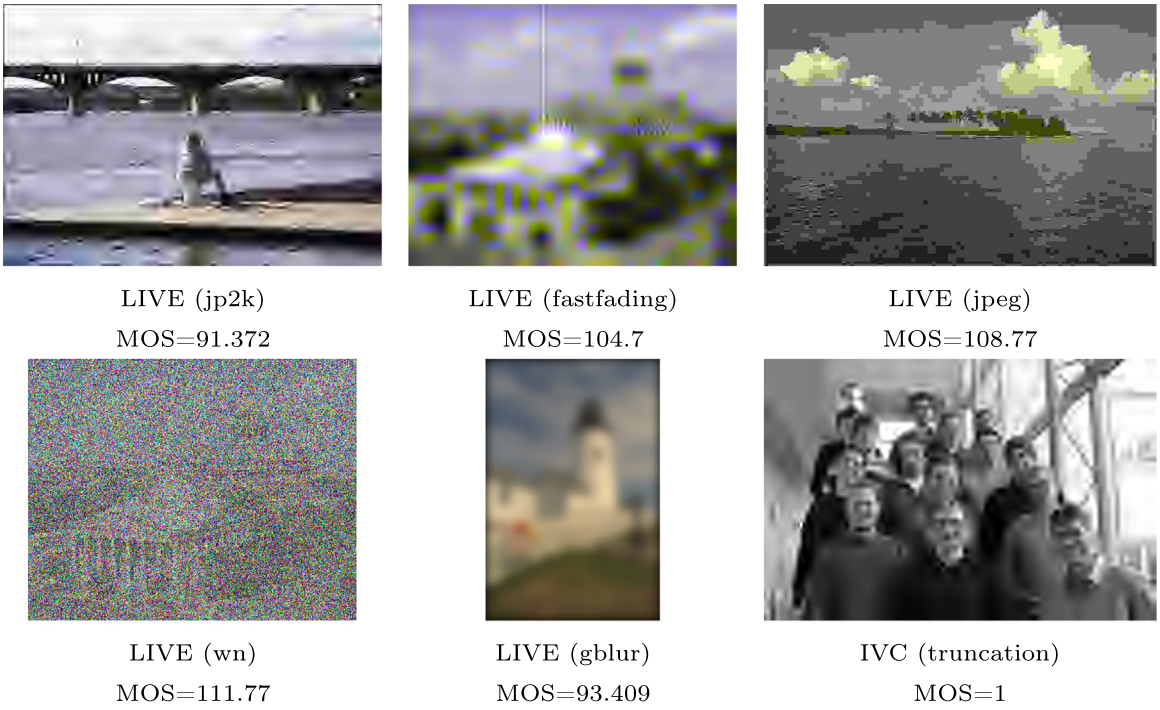


Fig. 12. The lowest quality images of each test set in the LIVE database as well as one of the lowest quality images from the IVC-SelectEncrypt database.

without losing too much information. The result is that the measures μ and σ combined with the signal shape actually contain a lot of information. The purpose of the detailed discussion in this section is to relate to the reader the implications carried by the 'scores' μ, σ and signal shape.

Fig. 13 shows the detailed evaluation of confidence on the LIVE database. For each metric, the figure shows a scatter plot of MOS and metric values, the bounding curves $V_{\min}(D)$ and $V_{\max}(D)$ as well as a plot of the local confidence value C_D. Table 2 lists the confidence scores $\mu(C_D)$ and $\sigma(C_D)$ for each metric on the LIVE database. In order to make the confidence scores comparable a pseudo normalization was used. Bounded metrics are normalized and unbounded metrics, e.g., PSNR, are normalized by mapping the range of occurring metric scores into [0, 1]. For the calculation of the signal shape the leading and trailing 10% of the MOS range where not taken into account because, at the high and low ends of the MOS, the difference in metric scores is limited due the boundary of the metric range, see Fig. 13. This would result in false outliers. The shape was calculated, similar to $\mu(C_D)$ and $\sigma(C_D)$, on pseudo normalized local confidence value C_D. In Fig. 13, outliers are indicated by down and up arrows for high and low quality outliers, respectively.

For the IVC-SelectEncrypt database the accordant plot and confidence scores are given in Fig. 14 and Table 3, respectively. Notice that while a MOS score of 0 is high quality on the LIVE database, a MOS score of 0 represents low quality on the IVC-SelectEncrypt database.

When it comes to confidence we can safely state that none of the metrics show good performance overall, i.e., none of the metrics are stable with a low $\mu(C_D)$ and $\sigma(C_D)$. However, for certain test sets there are metrics with good performance, e.g., CPA1 on the IVC-SelectEncrypt test set shows exceptionally high confidence and is stable. The CPA1 metric on the LIVE database, however, shows extremely poor performance.

Let us consider the relation between shape, $\mu(C_D)$ and $\sigma(C_D)$. The notional average performance is given by $\mu(C_D)$. The amount of deviation from the notional average is given by $\sigma(C_D)$. The way the deviation is distributed is given by the signal shape. Contrariwise, $\sigma(C_D)$ indicates the magnitude of the shape, i.e., assuming constant $\mu(C_D)$ a biased signal will be biased to a higher degree if $\sigma(C_D)$ is higher.

Compare LEG and SSIM as an example: While LEG shows better overall $\mu(C_D)$ and $\sigma(C_D)$, SSIM is much more biased than LEG, i.e., SSIM outperforms LEG where its bias is. Contrary to that, LEG outperforms SSIM outside of the bias. This behavior is only identifiable when $\sigma(C_D)$ is considered in conjunction with shape. In particular, even though the LEG is biased, due to the small $\sigma(C_D)$ we can deduce that the bias is far smaller than the bias of the SSIM (which exhibits a high $\sigma(C_D)$).

A similar behavior can be observed for unstable shapes. A metric with an unstable shape and a high $\sigma(C_D)$ will have much more severe outliers than one with a low $\sigma(C_D)$. This is nicely reflected in the PSNR: Although unstable, the magnitude of the outliers should be relatively small since the PSNR shows a low $\sigma(C_D)$. This is exactly what we see in the plots of Figs. 13 and 14.

For a stable shape, $\mu(C_D)$ becomes more important since it shows where the stable part of the confidence lies. The LE metric on the IVC-SelectEncrypt database is a prime example of this: While it is stable, the actual confidence score shows that it is stable in the sense that it exhibits poor performance over the whole quality range.

Regarding confidence values and shape, it can be seen from Tables 2 and 3 that can exhibit non-uniform behavior over different test sets. If this is the case, when calculating the overall performance scores, the worst value should be taken into account.

What is also noticeable from the two tables is the fact that the evaluated image and security metrics are more often biased towards the high quality range. Indeed, on the IVC-SelectEncrypt database which is the actual encryption database, not a single metric is biased towards the low quality range. Furthermore, the metrics biased towards the low quality range on the LIVE database, i.e., LEG, VIF and LFBVS, are all biased towards high quality on the IVC-SelectEncrypt database, and should thus be considered unstable overall.

To sum up the findings regarding the confidence of the metrics we can state the following: *First*, LE, LSS and CPA1 show extremely

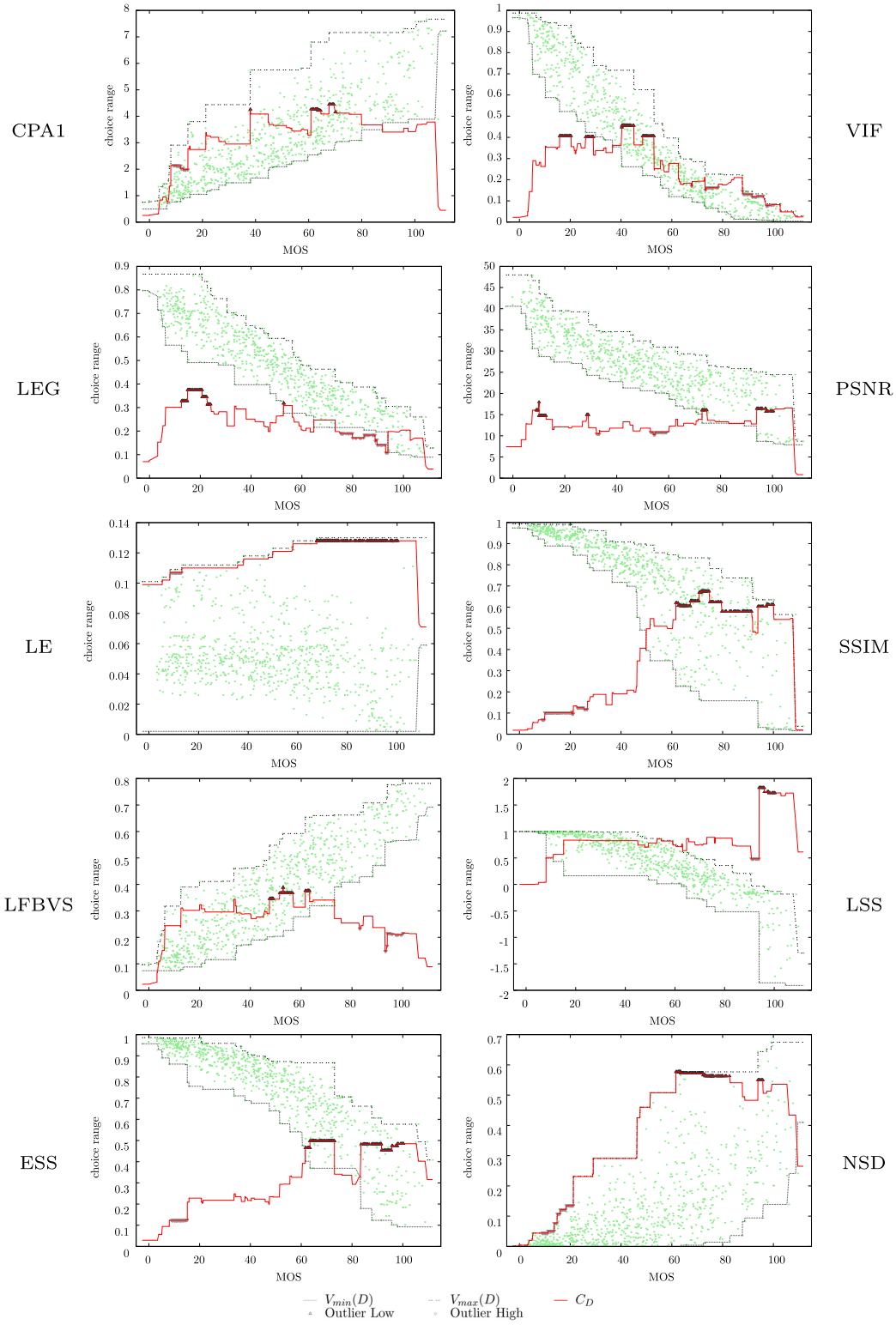


Fig. 13. Confidence plots for the LIVE database for different image metrics. The plot shows the scatter plot for the MOS and metric score pairs, the plot of $V_{min}(D)$, $V_{max}(D)$ and C_D.

Table 2
Average and standard deviation of normalized confidence and signal shape on the LIVE database.

	SSIM	LEG	VIF	CPA1	LSS	ESS	LFBVS	LE	NSD	PSNR
$\mu(C_D)$	0.357	0.291	0.285	0.431	0.415	0.300	0.370	0.906	0.537	0.265
$\sigma(C_D)$	0.225	0.070	0.110	0.109	0.159	0.133	0.071	0.069	0.277	0.038
Signal shape	Bias high	Bias low	Bias low	Bias high	Bias high	Bias high	Bias low	Bias high	Bias high	Unstable

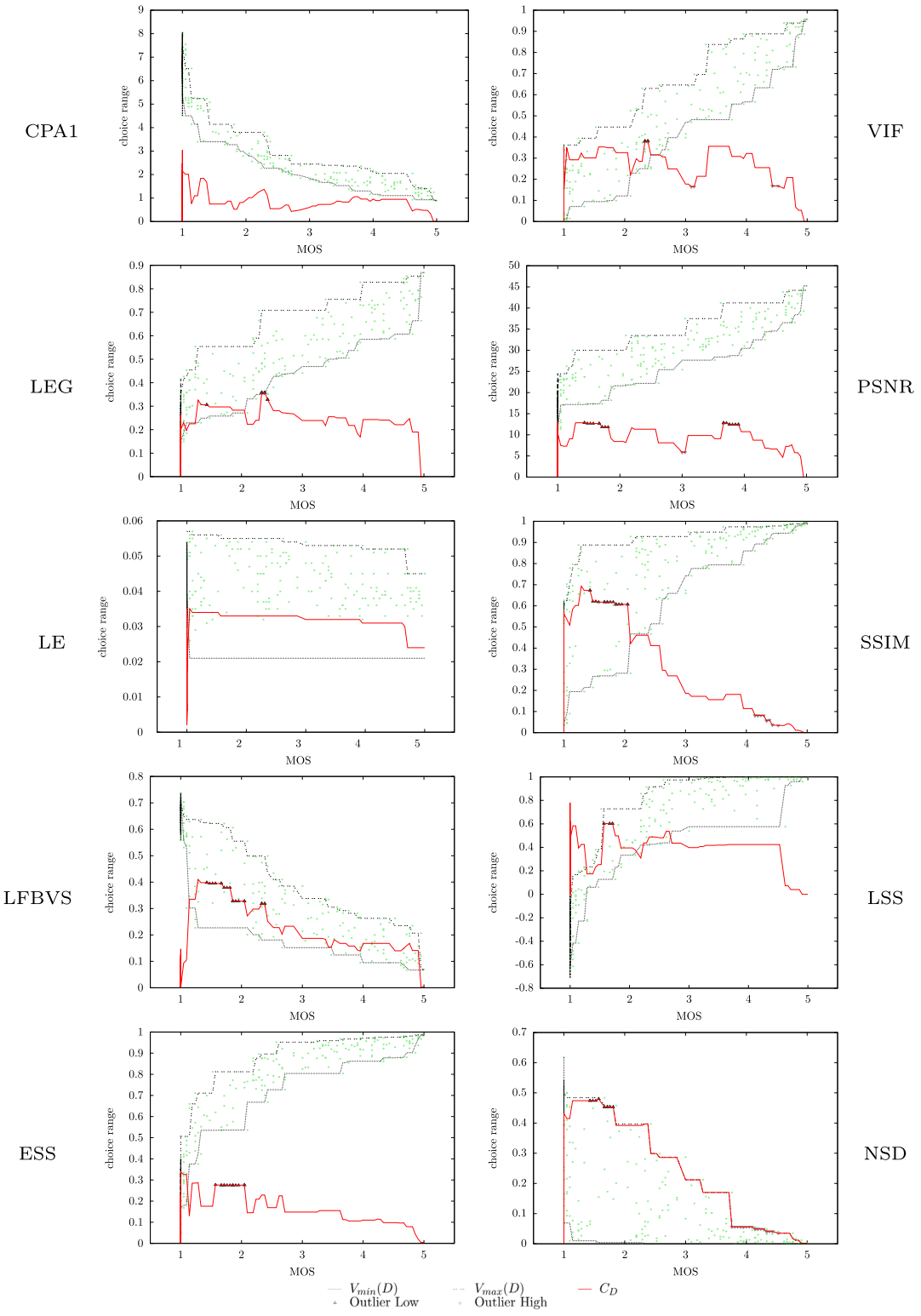


Fig. 14. Confidence plots for the IVC-SelectEncrypt database for different image metrics. The plot shows the scatter plot for the MOS and metric score pairs, the plot of $V_{min}(D)$, $V_{max}(D)$ and C_D.

Table 3
Average and standard deviation of normalized confidence and signal shape on the IVC-SelectEncrypt database.

	SSIM	LEG	VIF	CPA1	LSS	ESS	LFBVS	LE	NSD	PSNR
$\mu(C_D)$	0.319	0.268	0.277	0.119	0.374	0.168	0.273	0.540	0.394	0.196
$\sigma(C_D)$	0.226	0.077	0.098	0.066	0.173	0.090	0.139	0.107	0.274	0.063
Signal shape	Bias high	Bias high	Bias high	Stable	Bias high	Bias high	Bias high	Stable	Bias high	Unstable

poor performance overall. While CPA1 performs exceptionally well on the IVC-SelectEncrypt database, it performs poorly on the LIVE database; consequently, the overall performance of CPA1 is not good. *Second*, LEG, VIF, ESS, LFBVS and PSNR – while not stable – exhibit a low $\sigma(C_D)$ and are thus closest to being considered good metrics over the full quality range. However, in each case the $\mu(C_D)$ is relatively high overall (at least from a security standpoint) and thus could use some improvement, or replacement. *Third*, SSIM and NSD show a strong bias towards the high quality range. While the confidence for these metrics is not good overall, the confidence over the high quality range is actually quite good. Consequently, when the application scenario is known to target the higher quality range, e.g., in transparent encryption, these metrics should be considered.

4.3. Monotonicity for low quality images

The monotonicity of an image metric over the MOS for the full quality range is what defines the quality of an image metric. However, for transparent/sufficient encryption it is more important that a metric has good monotonicity properties on the lower quality range than on the full quality range. It is well known, c.f. [13], that this can be a problem so it is necessary to study this in more detail.

The absolute SROC values for the test sets of the LIVE database are given in Table 4(a) for the full quality range, the low quality range and the high quality range. In the table, *high* means SROC >0.9 (marked bold); a SROC score of <0.5 means *unsatisfactory* (underlined). For the IVC-SelectEncrypt database the same information is listed in Table 4(b).

In order to better compare the difference in high, low and overall, rank order correlation for a given test set and database is illustrated in Table 5 for the LIVE and IVC-SelectEncrypt database. For each combination of test set and metric, the graphical entry displays the range of possible SROC scores as background, with SROC=0 at the bottom and SROC=1 at the top. The light gray background bar shows the SROC value for the full quality range while the smaller bars indicate the SROC score for the high quality (left) and low quality range (right). What can be directly seen is that the overall performance of a metric does not imply good performance for either the high or low quality range, although most metrics tend to perform better on the high quality range. In some cases, a metric can even perform better on a limited quality range than over the whole range of quality. For example, the VIF performs better for the low quality end of resolution encryption on the IVC-SelectEncrypt database than for the whole range of resolution. In other cases, the performance over the full quality range is drastically reduced over the high and low parts of the quality range. For example, the VIF performs poorly on the high quality range for resolution encryption on the IVC-SelectEncrypt database. For other test sets, the impact of either low or high quality is minor and most likely due to the reduced number of samples from the database, e.g., the VIF performs well for white noise distortion (from the LIVE database) irrespective of the quality range. This in essence shows that the actual performance of an image metric is dependent on the distortion type as well as the quality range. For most test sets and image metric combinations, the SROC over the full quality range can at most be used as an upper bound for the limited quality cases. However, there are exceptions, e.g., VIF for the low quality range of the resolution test set performs better than on the overall quality range for the same test set.

With regard to security metrics and their performance on the low quality range, there is no noticeable difference in behavior to regular image metrics. Interestingly, though, for each test set the best performance over the low quality range can be found among

Table 4
Spearman's rank order correlation (SROC) for full, high and low quality range.

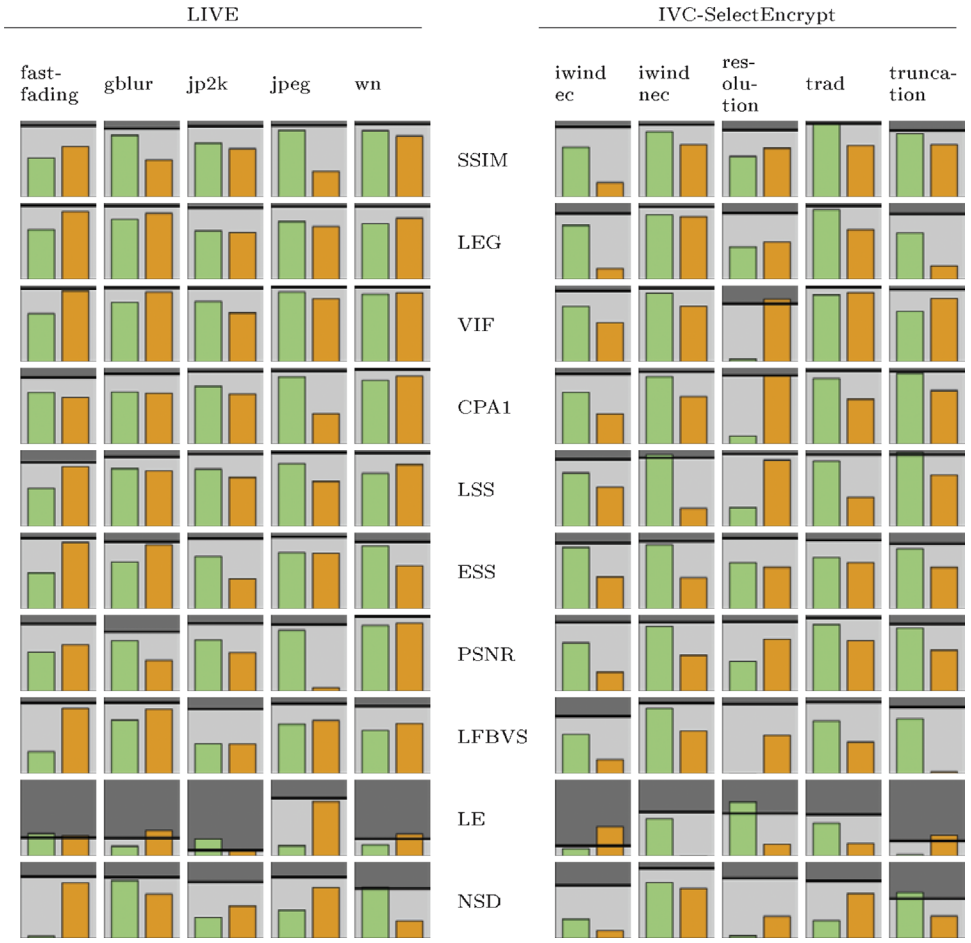
(a) SROC for the LIVE database

	fast-fading	gblur	jp2k	jpeg	wn
Full quality range					
SSIM	**0.942**	**0.903**	**0.936**	**0.946**	**0.962**
LEG	**0.971**	**0.966**	**0.945**	**0.960**	**0.960**
VIF	**0.965**	**0.972**	**0.968**	**0.984**	**0.985**
CPA1	0.881	**0.927**	**0.958**	**0.962**	**0.984**
LSS	0.843	**0.916**	**0.953**	**0.970**	**0.965**
ESS	**0.933**	0.888	**0.929**	**0.949**	0.886
PSNR	0.891	0.782	0.895	0.881	**0.985**
LFBVS	**0.932**	**0.937**	0.853	**0.920**	0.891
LE	0.241	0.236	0.076	0.766	0.228
NSD	0.815	0.803	0.741	0.806	0.657
High quality range					
SSIM	0.517	0.815	0.710	0.879	0.875
LEG	0.652	0.792	0.638	0.761	0.734
VIF	0.632	0.782	0.797	**0.920**	0.891
CPA1	0.681	0.689	0.761	0.889	0.840
LSS	0.504	0.762	0.754	0.825	0.700
ESS	0.473	0.618	0.694	0.744	0.827
PSNR	0.517	0.662	0.676	0.801	0.869
LFBVS	0.289	0.707	0.393	0.651	0.572
LE	0.295	0.127	0.228	0.136	0.143
NSD	0.028	0.760	0.275	0.370	0.674
Low quality range					
SSIM	0.662	0.487	0.635	0.339	0.802
LEG	0.893	0.872	0.617	0.699	0.804
VIF	**0.937**	**0.920**	0.646	0.829	**0.911**
CPA1	0.614	0.669	0.657	0.402	0.897
LSS	0.788	0.732	0.647	0.595	0.817
ESS	0.877	0.847	0.397	0.735	0.570
PSNR	0.611	0.408	0.510	0.046	0.897
LFBVS	0.863	0.849	0.391	0.702	0.659
LE	0.268	0.338	0.070	0.720	0.290
NSD	0.731	0.582	0.427	0.668	0.231

(b) SROC for the IVC-SelectEncryptdatabase

	iwindec	iwindnec	resolution	tradition	truncation
Full quality range					
SSIM	**0.925**	**0.954**	0.887	**0.968**	0.879
LEG	0.869	**0.956**	0.876	**0.966**	0.863
VIF	**0.937**	**0.969**	0.767	**0.982**	**0.954**
CPA1	**0.925**	**0.953**	**0.906**	**0.967**	**0.959**
LSS	0.888	**0.907**	**0.955**	**0.948**	**0.948**
ESS	0.868	0.894	**0.933**	**0.906**	0.860
PSNR	**0.909**	**0.910**	**0.916**	**0.965**	0.889
LFBVS	0.756	**0.930**	**0.916**	**0.943**	0.878
LE	0.136	0.579	0.562	0.550	0.201
NSD	0.698	**0.922**	0.790	0.757	0.523
High quality range					
SSIM	0.652	0.863	0.714	**0.975**	0.835
LEG	0.713	0.852	0.643	**0.920**	0.610
VIF	0.729	**0.907**	0.143	0.885	0.662
CPA1	0.683	0.890	0.000	0.868	**0.934**
LSS	0.705	**0.945**	0.393	0.865	**0.975**
ESS	0.809	0.846	0.571	0.679	0.794
PSNR	0.636	0.857	0.500	0.879	0.830
LFBVS	0.521	0.863	0.250	0.698	0.723
LE	0.095	0.489	0.464	0.431	0.019
NSD	0.251	0.736	0.179	0.236	0.602
Low quality range					
SSIM	0.191	0.694	0.644	0.680	0.695
LEG	0.141	0.823	0.490	0.652	0.181
VIF	0.518	0.732	0.823	**0.913**	0.832
CPA1	0.400	0.628	0.897	0.592	0.706
LSS	0.523	0.240	0.875	0.386	0.679
ESS	0.422	0.411	0.551	0.609	0.549
PSNR	0.251	0.474	0.688	0.663	0.541
LFBVS	0.188	0.562	0.507	0.416	0.022
LE	0.386	0.004	0.152	0.166	0.272
NSD	0.100	0.655	0.290	0.589	0.291

Table 5
Visual representation of Spearman's rank order correlation ($|SROC| \in [0, 1]$) for the LIVE Image Quality Assessment and IVC-SelectEncrypt databases for full quality range (light gray), low quality range (orange bar on the right) and high quality range (green bar on the left side).



the traditional image metrics.

Comparing the high and low quality ranges, we can see that the low quality range has a far higher number of unsatisfactory SROC scores. This shows that image metrics tend to perform better at differentiating the different strength in distortion for high quality images. A high SROC score over the whole quality range indicates that the metric can differentiate between high and low quality images. In essence, image metrics which perform well on the overall quality range can still be utilized to identify sufficient encryption, even though a metric which exhibits good performance in the range of the quality threshold should be preferred. For transparent encryption, where the goal is to find the best image below a certain threshold, monotonicity in the chosen quality range becomes more important. In particular, the target quality is a lot closer to the threshold, thus a high monotonicity (expressed by a high SROC) is required.

Another interesting aspect of the high versus low quality test is the fact that none of the tested metrics exhibits a better performance on the high or low quality range for all test sets. Consequently, the metrics cannot be reduced to a single SROC score and a bias towards either high or low. There are cases where the performance over both high and low quality is far worse than for the whole quality range, e.g., LFBVS on the jp2k test set. Thus, in order to evaluate whether an image metric can be used as a security metric, tests regarding low and high quality performance have to be conducted.

Summing up the monotonicity tests, we can state the following: *First*, VIF, CPA1, SSIM, LSS and LEG perform best over the whole quality range. LFBVS, PSNR and ESS also show a good behavior, while NSD and especially LE perform poorly.

Second, for the high quality range, most quality metrics still show a decent performance. However, only SSIM, LEG, and PSNR exhibit no unsatisfactory performance in a single test set.

Third, for the low quality range, only VIF exhibits good performance over all test sets. All other metrics have at least two test sets where their performance is unsatisfactory.

5. Conclusion and future work

We have outlined an evaluation method for security metrics based on practical application scenarios and considerations. These methods were used to evaluate (1) state-of-the-art security metrics, (2) image metrics which are used as security metrics as well as (3) state-of-the-art image metrics. A summary of the evaluation and a basic evaluation score is given in Table 6. To simplify and give a clear comparison between metrics, we assign a single score to each metric based on its performance in the various evaluation steps. In particular, for each step, we assign a score of 1 for desired behavior and a penalty of -1 for a clear failure. The final score is

Table 6
Summary of evaluation.

	SSIM	LEG	VIF	CPA1	LSS	ESS	PSNR	LFBVS	LE	NSD
Application domain										
Encryption	*0.450*	0.334	*0.454*	0.363	*0.417*	0.373	*0.512*	*0.520*	0.731	*0.460*
Extraction	**0.999**	**0.994**	**0.988**	**0.042**	0.692	**0.958**	0.681	*0.489*	*0.500*	*0.438*
Confidence on the LIVE database										
$\mu(C_D)$	0.357	**0.291**	**0.285**	0.431	0.415	0.300	**0.265**	0.370	*0.906*	*0.537*
$\sigma(C_D)$	0.225	**0.070**	0.110	0.109	0.159	0.133	**0.038**	**0.071**	**0.069**	*0.277*
Signal shape	Bias high	Bias low	Bias low	Bias high	Bias high	Bias high	Bias low	Bias high	Bias high	*Unstable*
Confidence on the IVC-SelectEncrypt database										
$\mu(C_D)$	0.319	**0.268**	**0.277**	**0.119**	0.374	**0.168**	**0.196**	**0.273**	*0.540*	0.394
$\sigma(C_D)$	0.226	**0.077**	**0.098**	**0.066**	0.173	**0.090**	**0.063**	0.139	0.107	*0.274*
Signal shape	Bias high	Bias high	Bias high	Stable	Bias high	Bias high	Bias high	Stable	Bias high	*Unstable*
Low quality SROC on the LIVE database										
fastfading	0.662	0.893	**0.937**	0.614	0.788	0.877	0.611	0.863	*0.268*	0.731
gblur	*0.487*	0.872	**0.920**	0.669	0.732	0.847	*0.408*	0.849	*0.338*	0.582
jp2k	0.635	0.617	0.646	0.657	0.647	*0.397*	0.510	*0.391*	*0.070*	*0.427*
jpeg	*0.339*	0.699	0.829	*0.402*	0.595	0.735	*0.046*	0.702	0.720	0.668
wn	0.802	0.804	**0.911**	0.897	0.817	0.570	0.897	0.659	*0.290*	*0.231*
Low quality SROC on the IVC-SelectEncrypt database										
iwind ec	*0.191*	*0.141*	0.518	*0.400*	0.523	*0.422*	*0.251*	*0.188*	*0.386*	*0.100*
iwind nec	0.694	0.823	0.732	0.628	*0.240*	*0.411*	*0.474*	0.562	*0.004*	0.655
resolution	0.644	*0.490*	0.823	0.897	0.875	0.551	0.688	0.507	*0.152*	*0.290*
trad	0.680	0.652	**0.913**	0.592	*0.386*	0.609	0.663	*0.416*	*0.166*	0.589
truncation	0.695	*0.181*	0.832	0.706	0.679	0.549	0.541	*0.022*	*0.272*	*0.291*
Comparison score, -1 or $+1$ for insufficient or good performance, -1 for conflict in signal shape										
Score	*−3*	1	**6**	0	*−3*	0	*−2*	*−5*	*−11*	*−12*

the sum of the individual test scores. In Table 6, desired behavior is indicated in bold and failure in italics.

For the *application domain* a sorting error of less than 10% is considered good, while a performance around 50% ($\pm 10\%$) is considered a failure. For the *confidence*, a score for $\mu \pm \sigma$ of 0.3 ± 0.1 is considered good, while a result of 0.5 ± 0.25 is considered a failure. Regarding signal shape, we desire stable signals. The signal shape will also be penalized if it does not agree on the two test sets. For the *quality and SROC*, a value ≤ 0.5 is a failure while an SROC greater than 0.9 is desirable. From the final scores in Table 6, we conclude that none of the security metrics are fit to perform as general purpose security metrics.

Regarding transparent and sufficient encryption, the LE and NSD metrics especially show that metrics engineered to fit a certain application scenario cannot claim generality. Furthermore, SSIM and PSNR, which are frequently used as security metrics, also perform poorly. Most state-of-the-art image metrics hardly perform the security metric task adequately. Only VIF, apart from a borderline confidence score and stability, demonstrates good performance.

Regarding content confidentiality we cannot make a strong statement due to lack of ground truth for recognizability tests. However, the performance in the encrypted domain, during the evaluation of the application domain, gives a strong indication that none of the tested image metrics can perform the task of evaluating the content confidentiality.

The presented methodology for the evaluation of security metrics should be used to asses newly developed metrics to prevent similar shortcomings as those shown by the investigated metrics. Table 6 gives a concise overview of the state-of-the-art metrics, and can be used as a basis for comparison for new image and security metrics.

5.1. Future work

The inability to properly evaluate image metrics with regard to content confidentiality naturally leads to the conclusion that more data is required. This also holds true for the lower quality ranges with respect to regular metrics, which would undoubtedly benefit from a dataset that is specifically designed for high impairment cases. In future work, we will gather ground truth data for content confidentiality (and will also design protocols how to properly capture human assessment for these data sets) and extend the work presented in this paper to properly encompass content confidentiality. Furthermore, we intend to gather more data on the low quality range as well, in order to better evaluate security metrics for transparent and sufficient encryption.

Acknowledgments

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

References

[1] S.-K. Au Yeung, S. Zhu, B. Zeng, Quality assessment for a perceptual video encryption system, in: 2010 IEEE International Conference on Wireless Communications, Networking and Information Security (WCNIS), 2010, pp. 102–106, http://dx.doi.org/10.1109/WCINS.2010.5541898.
[2] M. Khan, V. Jeoti, A. Malik, Perceptual encryption of JPEG compressed images using DCT coefficients and splitting of DC coefficients into bitplanes, in: 2010 International Conference on Intelligent and Advanced Systems (ICIAS), 2010, pp. 1–6.
[3] E.J. Wharton, K. Panetta, S. Agaian, Simultaneous encryption/compression of images using alpha rooting, in: 2008 Data Compression Conference (DCC 2008), 25–27 March 2008, Snowbird, UT, USA, IEEE Computer Society, 2008, p. 551ff, https://www.computer.org/csdl/proceedings/dcc/2008/3121/00/3121a551-abs.html.
[4] S. Lian, Efficient image or video encryption based on spatiotemporal chaos system, Chaos Solitons Fractals 40 (5) (2009) 2509–2519, http://dx.doi.org/10.1016/j.chaos.2007.10.054, URL ⟨http://www.sciencedirect.com/science/article/pii/S0960077907009277⟩.
[5] C.E. Shannon, Communication theory of secrecy systems, Bell Syst. Tech. J. 28 (1949) 656–715.
[6] T.D. Lookabaugh, D.C. Sicker, Selective encryption for consumer applications, IEEE Commun. Mag. 42 (5) (2004) 124–129.
[7] A. Said, Measuring the strength of partial encryption schemes, in: Proceedings

of the IEEE International Conference on Image Processing (ICIP'05), vol. 2, 2005, pp. 1126–1129, http://dx.doi.org/10.1109/ICIP.2005.1530258.
[8] M. Bellare, T. Ristenpart, P. Rogaway, T. Stegers, Format-preserving encryption, in: Proceedings of Selected Areas in Cryptography, SAC '09, vol. 5867, Springer-Verlag, Calgary, Canada, 2009, pp. 295–312.
[9] T. Stütz, A. Uhl, A survey of H.264 AVC/SVC encryption, IEEE Trans. Circuits Syst. Video Technol. 22 (3) (2012) 325–339.
[10] T. Stütz, A. Uhl, Efficient format-compliant encryption of regular languages: block-based cycle-walking, in: B.D. Decker, I. Schaumüller-Bichl (Eds.), Proceedings of the 11th Joint IFIP TC6 and TC11 Conference on Communications and Multimedia Security, CMS '10, IFIP Advances in Information and Communication Technology, vol. 6109, Springer, Linz, Austria, 2010, pp. 81–92.
[11] Q. Li, I.J. Cox, Using perceptual models to improve fidelity and provide resistance to valumetric scaling for quantization index modulation watermarking, IEEE Trans. Inf. Forensics Secur. 2 (2) (2007) 127–139.
[12] H. Hofbauer, A. Uhl, Selective encryption of the MC-EZBC bitstream and residual information, in: 18th European Signal Processing Conference, 2010 (EUSIPCO-2010), Aalborg, Denmark, 2010, pp. 2101–2105.
[13] H. Hofbauer, A. Uhl, Visual quality indices and low quality images, in: IEEE 2nd European Workshop on Visual Information Processing, Paris, France, 2010, pp. 171–176.
[14] Y. Zhou, K. Panetta, S. Aagaian, Partial multimedia encryption with different security levels, in: 2008 IEEE Conference on Technologies for Homeland Security, 2008, http://dx.doi.org/10.1.1.152.725, URL ⟨http://biron.usc.edu/~sagetong/resume.html⟩.
[15] B. SaiChandana, S. Anuradha, A new visual cryptography scheme for color images, Int. J. Eng. Sci. Technol. 2 (6) (2010) 1998–2000.
[16] L. Dubois, W. Puech, J. Blanc-Talon, Confidentiality metrics and smart selective encryption for HD H.264/AVC videos, in: 2013 Proceedings of the 21st European Signal Processing Conference (EUSIPCO), 2013, pp. 1–5.
[17] L. Dubois, W. Puech, J. Blanc-Talon, Smart selective encryption of cavlc for h.264/avc video, in: 2011 IEEE International Workshop on Information Forensics and Security (WIFS), 2011, pp. 1–6, http://dx.doi.org/10.1109/WIFS.2011.6123130.
[18] W. Wang, D. Peng, H. Wang, H. Sharif, An adaptive approach for image encryption and secure transmission over multirate wireless sensor networks, Wirel. Commun. Mob. Comput. 9 (3) (2009) 383–393, http://dx.doi.org/10.1002/wcm.550.
[19] Z. Shahid, W. Puech, Visual protection of HEVC video by selective encryption of CABAC binstrings, IEEE Trans. Multimed. 16 (1) (2014) 24–36, http://dx.doi.org/10.1109/TMM.2013.2281029.
[20] Q. Huynh-Thu, M. Ghanbari, Scope of validity of PSNR in image/video quality assessment, Electron. Lett. 44 (13) (2008) 800–801.
[21] F. Autrusseau, T. Stuetz, V. Pankajakshan, Subjective quality assessment of selective encryption techniques, ⟨http://www.irccyn.ec-nantes.fr/~autrusse/Databases/⟩, 2010.
[22] Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[23] Y. Mao, M. Wu, Security evaluation for communication-friendly encryption of multimedia, in: Proceedings of the IEEE International Conference on Image Processing (ICIP'04), IEEE Signal Processing Society, Singapore, 2004.
[24] H. Hofbauer, A. Uhl, Selective encryption of the MC EZBC bitstream for DRM scenarios, in: Proceedings of the 11th ACM Workshop on Multimedia and Security, ACM, Princeton, New Jersey, USA, 2009, pp. 161–170.
[25] H. Hofbauer, T. Stütz, A. Uhl, Selective encryption for hierarchical MPEG, in: H. Leitold, E. Markatos (Eds.), Communications and Multimedia Security, Proceedings of the 10th IFIP International CMS 2006 Conference, Lecture Notes on Computer Science, vol. 4237, Springer Verlag, Heraklion, Crete, 2006, pp. 151–160.
[26] Y. Yao, Z. Xu, J. Sun, Visual security assessment for cipher-images based on neighborhood similarity, Informatica 33 (2009) 69–76.
[27] J. Sun, Z. Xu, J. Liu, Y. Yao, An objective visual security assessment for cipher-images based on local entropy, Multimed. Tools Appl., 53 (1) (2011) 75–95, http://dx.doi.org/10.1007/s11042-010-0491-5.
[28] L. Tong, F. Dai, Y. Zhang, J. Li, Visual security evaluation for video encryption, in: Proceedings of the International Conference on Multimedia, MM '10, ACM, New York, NY, USA, 2010, pp. 835–838, http://dx.doi.org/10.1145/1873951.1874091.
[29] H. Hofbauer, A. Uhl, An effective and efficient visual quality index based on local edge gradients, in: IEEE 3rd European Workshop on Visual Information Processing, Paris, France, 2011, p. 6pp.
[30] H.R. Sheikh, A.C. Bovik, Image information and visual quality, IEEE Trans. Image Process. 15 (2) (2006) 430–444.
[31] M. Carosi, V. Pankajakshan, F. Autrusseau, Towards a simplified perceptual quality metric for watermarking applications, in: Proceedings of SPIE, Multimedia on Mobile Devices, vol. 7542, SPIE, San Jose, CA, USA, 2010.
[32] T. Stütz, A. Uhl, On JPEG2000 error concealment attacks, in: Advantages in Image and Video Technology: Proceedings of the 3rd Pacific-Rim Symposium on Image and Video Technology, PSIVT '09, Lecture Notes in Computer Science, Springer, Tokyo, Japan, 2009, pp. 851–861.
[33] C. Spearman, The proof and measurement of association between two things, Am. J. Psychol. 100 (3/4) (1904) 441–471.
[34] M.G. Kendall, A new measure of rank correlation, Biometrika 30 (1/2) (1938) 81–93.
[35] T. Stütz, V. Pankajakshan, F. Autrusseau, A. Uhl, H. Hofbauer, Subjective and objective quality assessment of transparently encrypted JPEG2000 images, in: Proceedings of the ACM Multimedia and Security Workshop (MMSEC '10), ACM, Rome, Italy, 2010, pp. 247–252.
[36] H.R. Sheikh, Z. Wang, L. Cormack, A.C. Bovik, LIVE image quality assessment database release 2, 2006 ⟨http://live.ece.utexas.edu/research/quality⟩.
[37] H.R. Sheikh, M.F. Sabir, A.C. Bovik, A statistical evaluation of recent full reference image quality assessment algorithms, IEEE Trans. Image Process. 15 (11) (2006) 3440–3451.