Applicability of No-Reference Visual Quality Indices for Visual Security Assessment

Heinz Hofbauer, Andreas Uhl

University of Salzburg
{hofbauer,uhl}@cs.sbg.ac.at

IH & MMSec 2018

Introduction—Selective Encryption

We deal with a partial or selective encryption.

What is that?

This means only a selected part of the plain text is encrypted.

Why do we want that?

- We might want to allow for a preview.
- Encrypting less takes less time.
- We want the result to still be the same medium (format compliance).

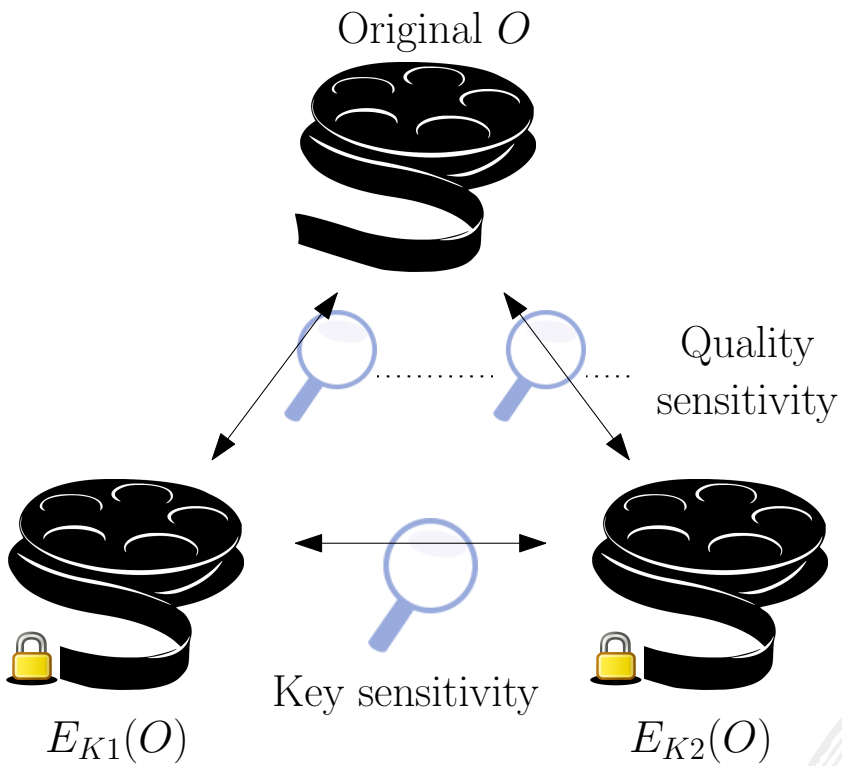
Introduction—Goals of Selective Encryption

Confidential Encryption where the content must not be recognizable
(and a recognizable reconstruction must not be possible.)

Sufficient Encryption where quality must be low
(and a higher quality reconstruction must not be possible.)

Transparent Encryption where quality must be near a defined value
(and a higher quality reconstruction must not be possible.)

Introduction—So why use Quality Indices?



Quality Sensitivity is the property that the quality of a cypher text should only be dependant on the selection (not on the key).

Quality Index Assessment (QIA)

Quality indices can be used for that type of sensitivity check, if they are fit for the purpose.

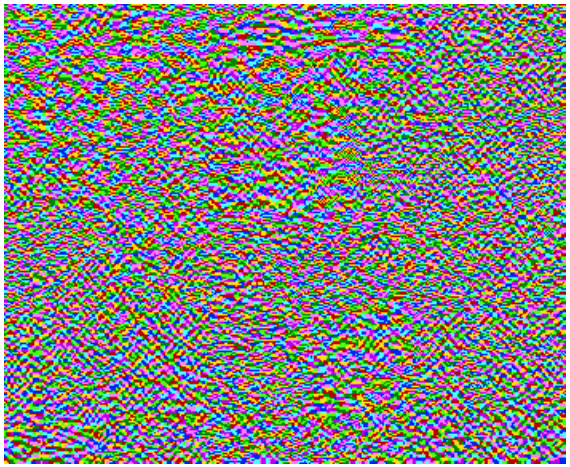
So how are the quality indices assessed themselves?

Asess the following:

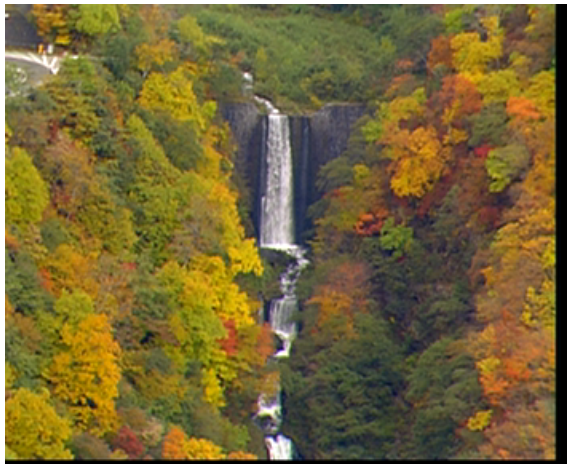
- In what **domain** can the index be apllied?
- **Correspondence** to human visual system:
  - **Monotonicity**
  - **Confidence**

QIA—Domain I

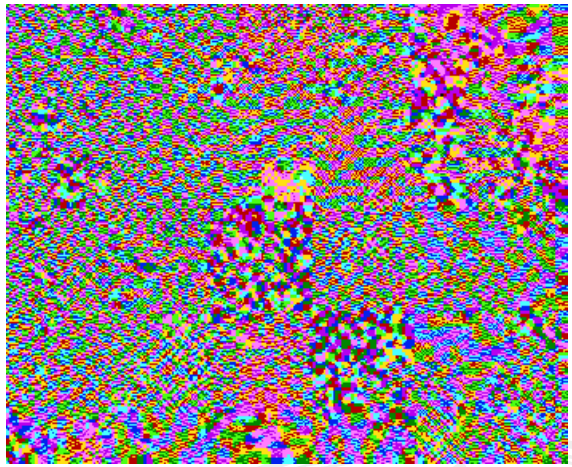
Guessing Game:



option 1



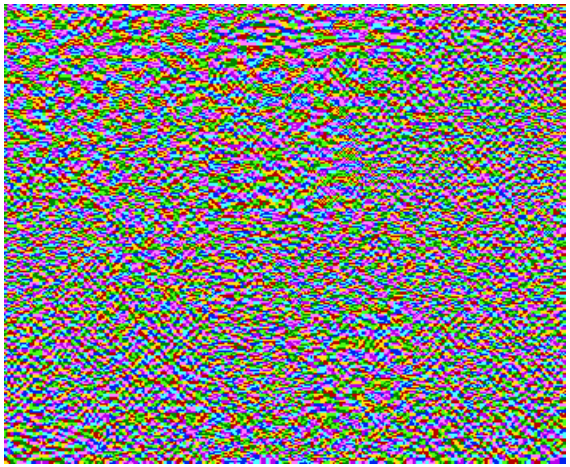
original



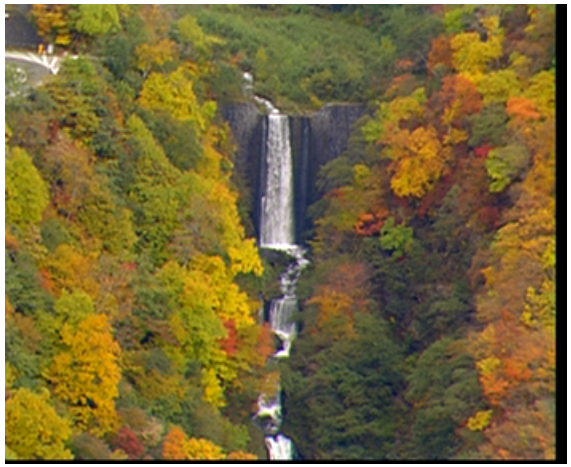
option 2

QIA—Domain I

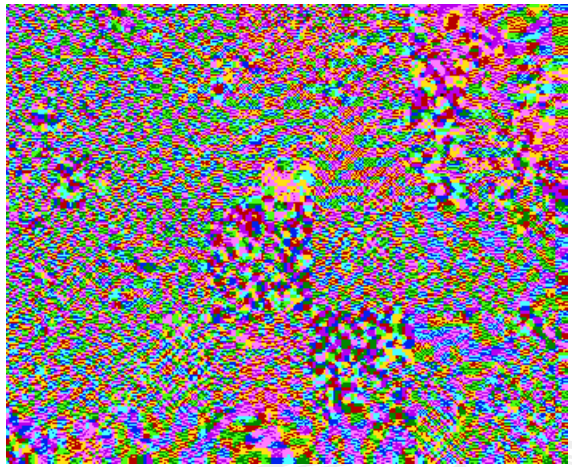
Guessing Game:



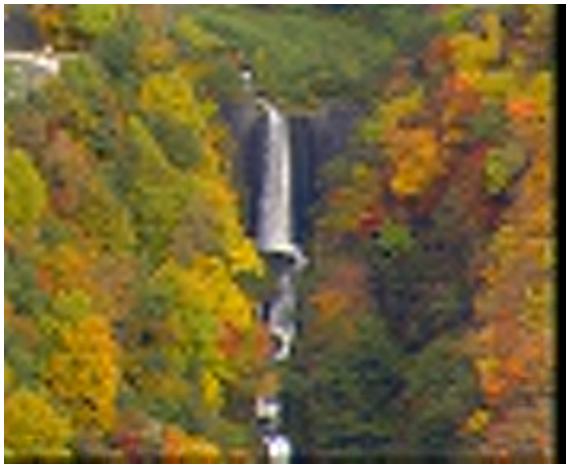
option 1

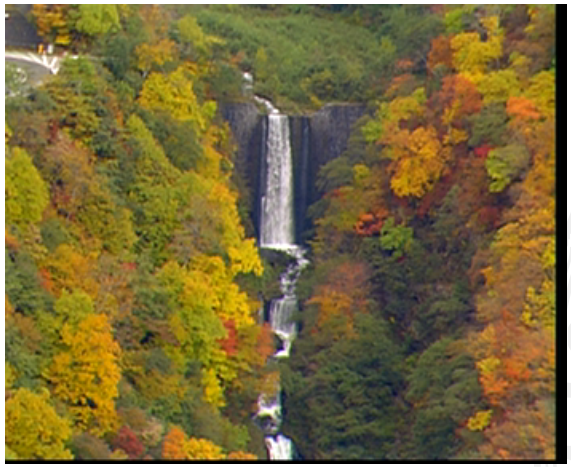


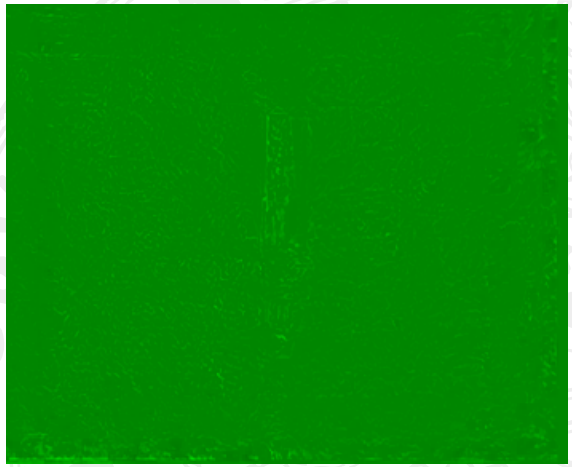
original



option 2

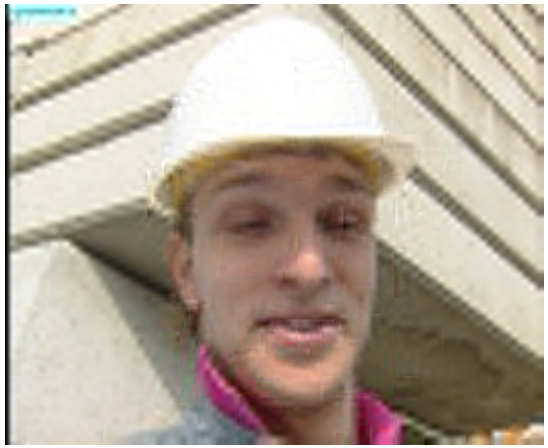


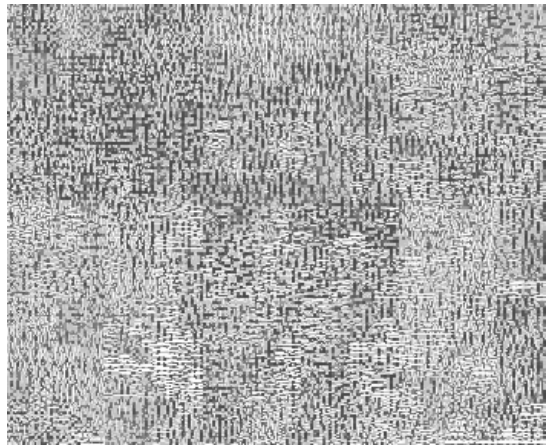




QIA—Domain II

Generate from an original two version of different quality:





The quality index should order them based on quality (often because statistics $\rightarrow$ N$\uparrow\uparrow$), and the higher the ratio of correctly ordered images the better the QI can be applied in the tested domain.

Do this for the

- Encrypted domain
- Extracted domain

QIA—Correspondence to the Human visual System

The quality indices are supposed to replace human observeration tests and should reflect human judgement.

This can be measured in two ways:

- Monotonicity: If a human observer would rate one image as higher quality than another then the quality index should do the same. The relation between human score and index score should be monotonous (linear would also be fine but is not really required).
- Confidence: Perfect monotonicity will not be achieved, even humans can disagree. Beauty is in the eye of the beholder" and all that. But we should have some indication that the disagreement is not too strong. Otherwise the metric can not be trusted.

QIA—Correspondence—Monotonicity

This basically the classical way of evaluating visual quality indices:

1. Take a database of human observer scores on distorted images
2. Apply the QI and get the scores.
3. Calculate how well the two scores correspond (usually with a rank order metric like Spearman rank order correlation).

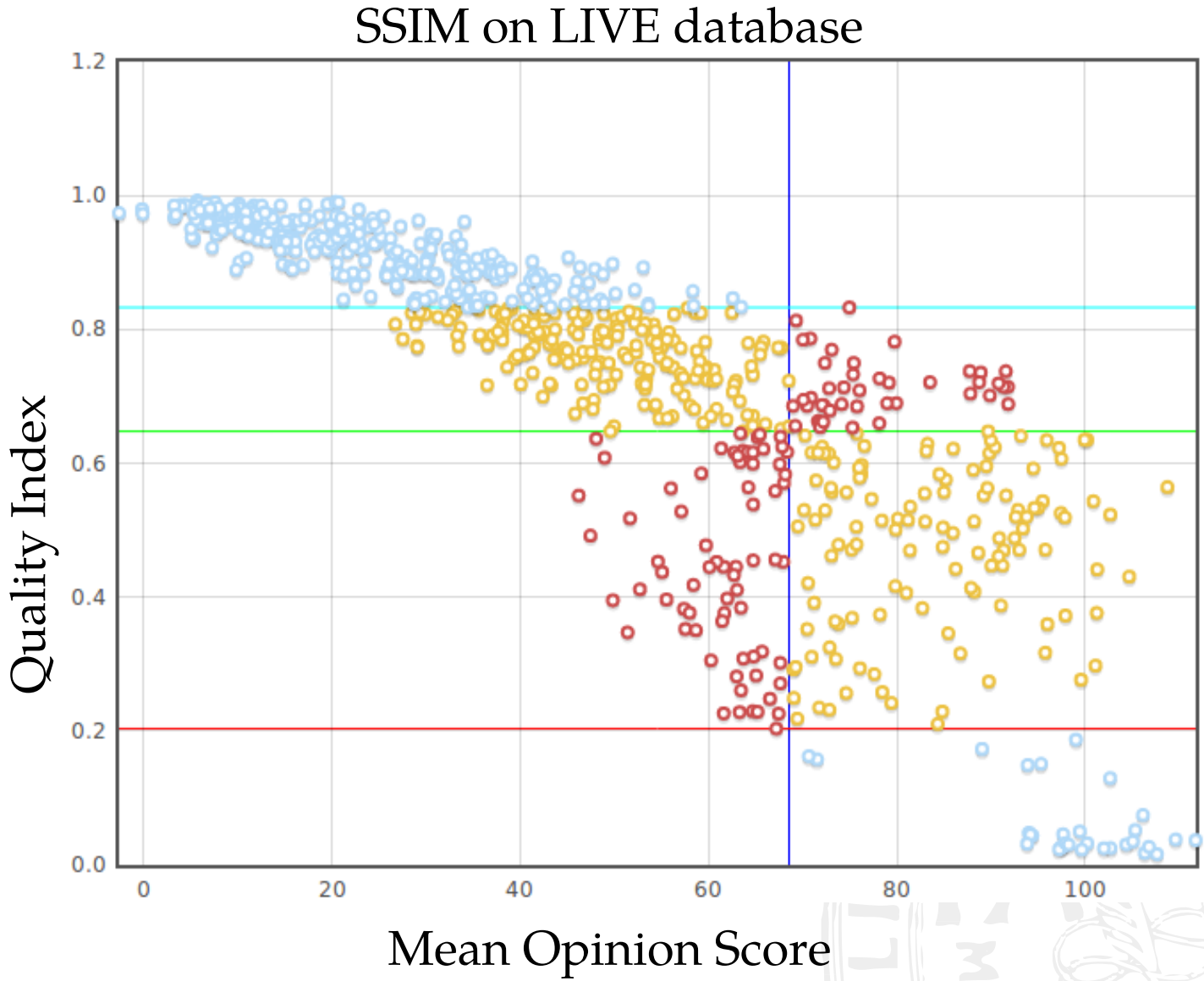
This can be done in a slightly more interesting way by separating the high and low quality images.

Why?

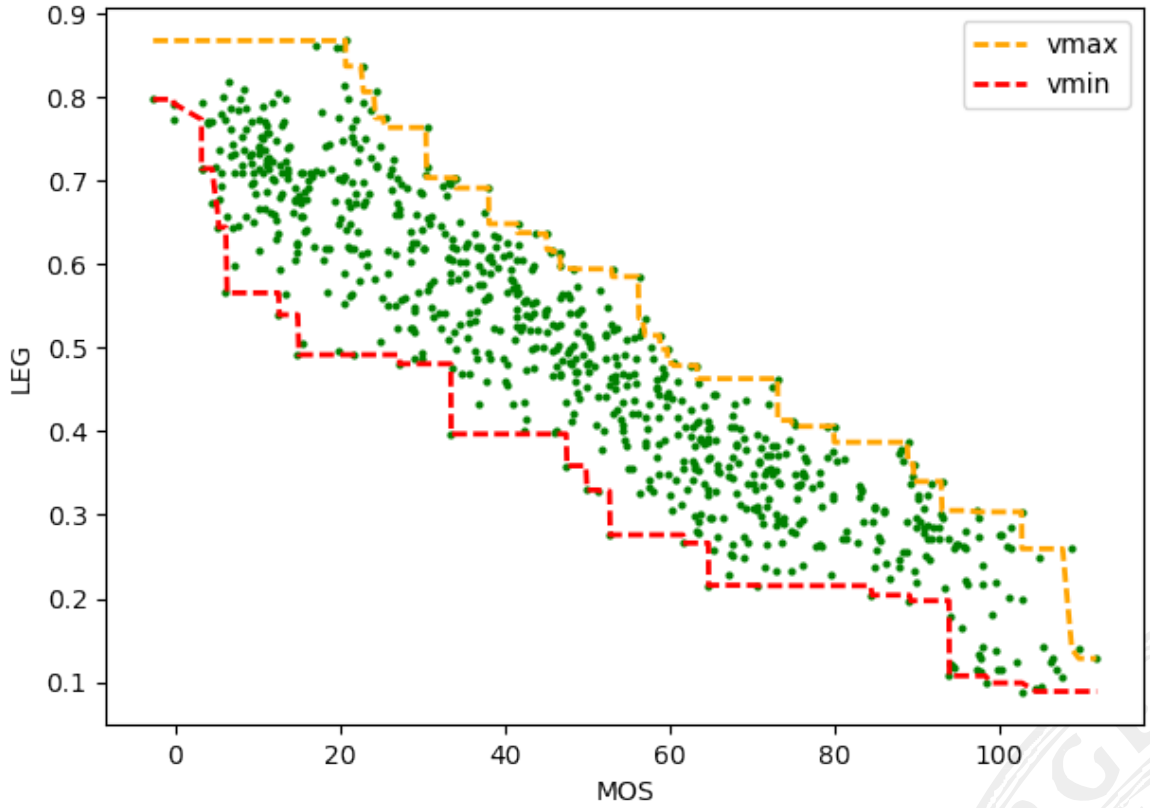
- Visual quality indices usually target a high quality range.
- For selective encryption we usually have a low quality range.

$\rightarrow$ the difference between these two shows us how the QI would work on it's intended subject as opposed to what we use it for.

QIA—Correspondence—Confidence I

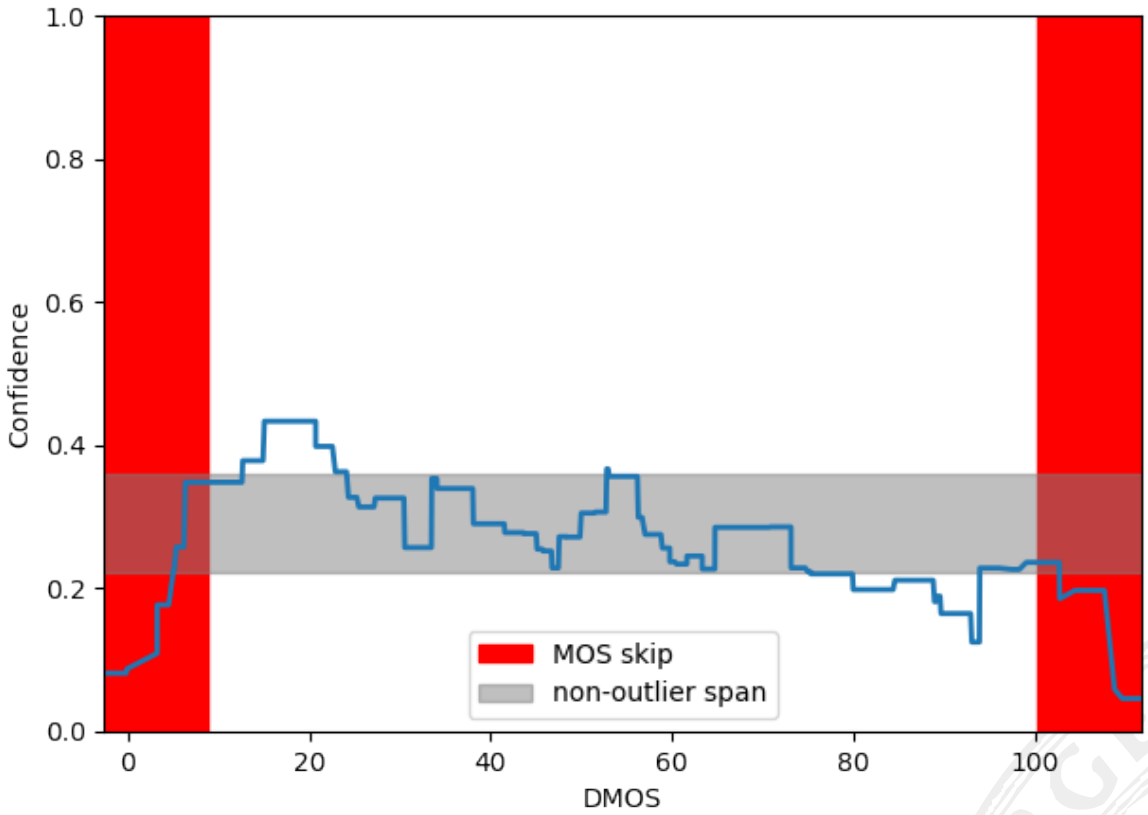


QIA—Correspondence—Confidence II



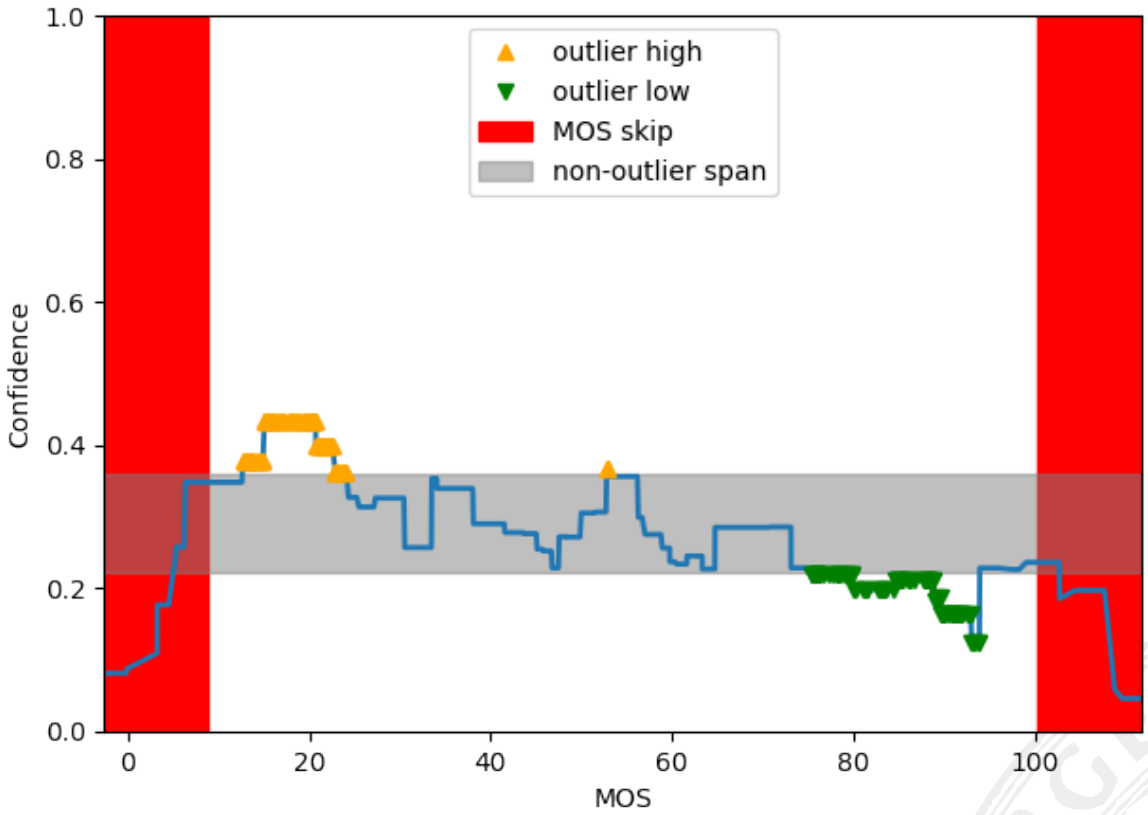
- $V_{min}(D)$ such that $\forall i : MOS(i) > D \implies v(i) > V_{min}(D)$ (zero false negatives)
- $V_{max}(D)$ such that $\forall i : v(i) > V_{max}(D) \implies MOS(i) > D$ (zero false positives)

QIA—Correspondence—Confidence III



- The confidence score $\mathcal{C}$ for a given MOS value D is $\mathcal{C}_D := |V_{max}(D) - V_{min}(D)|$.
- Outliers are simply calculated based on the z-score: $z(D, \mu, \sigma) = \frac{\mathcal{C}_D - \mu(\mathcal{C})}{\sigma(\mathcal{C})}$.

QIA—Correspondence—Confidence IV



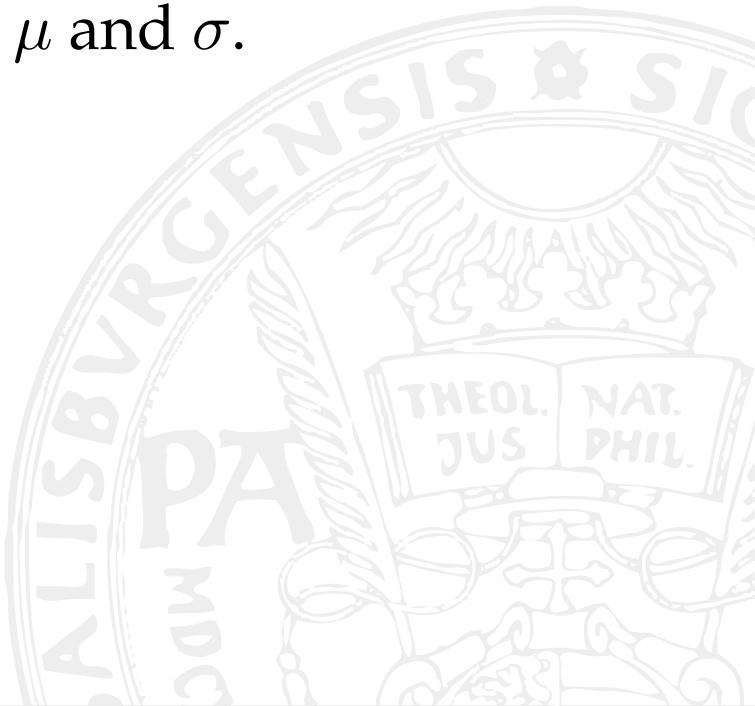
Can the outliers be separated or are they intermixed?

- **biased** if separable, further
  - biased high / low if low outlier at high / low quality
- **stable** if there are no outliers **unstable** otherwise

QIA—Correspondence—Confidence V

What do we want?

- Rank order correlation should be high
- Confidence Score should be low, both μ and σ.
- Stable signal (no outliers).



Blind Visual Quality Indices—Why? (I)

These are the best full reference quality indices which were tested in the past:

	LEG	VIF
Application Domain		
Encryption	0.334	0.454
Extraction	0.994	0.988
Confidence—LIVE DB		
$\mu(\mathcal{C}_D)$	0.291	0.285
$\sigma(\mathcal{C}_D)$	0.070	0.110
Signal Shape	Bias Low	Bias Low
Confidence—IVC DB		
$\mu(\mathcal{C}_D)$	0.268	0.277
$\sigma(\mathcal{C}_D)$	0.077	0.098
Signal Shape	Bias High	Bias High

	LEG	VIF
Low Quality SROC—LIVE DB		
fastfading	0.893	0.937
gblur	0.872	0.920
jp2k	0.617	0.646
jpeg	0.699	0.829
wn	0.804	0.911
Low Quality SROC—IVC DB		
iwind ec	0.141	0.518
iwind nec	0.823	0.732
resolution	0.490	0.823
trad	0.652	0.913
truncation	0.181	0.832
Comparison Score		
Score	1	6

Blind Visual Quality Indices—Why? (II)

Full-reference visual quality indices are built:

- Based on what we think is relevant for (high) quality.
- Are evaluated for high quality and are now fixed.

No-reference visual quality indices are built and trained:

- Based on what we think is relevant for (high) quality.
- They learn the actual distortions rather being engineered.
- Are not fixed, they can learn on low quality databases.

So the plan is as follows:

1. Evaluate NR-VQI as they are.
2. Try to learn them on low quality images (IVC SelectEncrypt database).

Blind Visual Quality Indices I

BIQAA: Pixel based directonal entropy.

BLIINDS-II: Generalized Gaussian distributions of features derived from groups of AC coefficients on multiple scales, including orientation.

BRISQUE: An asymmetric generalized Gaussian distribution (AGGD) is fitted to mean subtracted contrast normalized (MSNC) values (two scales, different directional differences).
Uses salable vector regression methods (SVR) to predict human observer scores instead of image statistics only.

Blind Visual Quality Indices II

Global phase coherence (GPC), Sharpness Index (sharp), and Simplified Index (SI): Difference of phase coherence (Fourier transform phase information) from a random signal. Sharp and SI use easier models to caculate for speed reasons.

NIQE: Directional MSNC values similar to BRISQUE@ but locally instead of globally.

SSEQ: Learns mean and skew of spatial and spectral entropies as features (multiple scales).
The features are trained using an SVR to conform to human judgement.

Evaluation—Application Domain

	VIF	BIQAA	BLIINDS-II	BRISQUE	GPC
Encryption	0.454	0.555	0.512	0.504	0.496
Extraction	0.988	0.535	0.000	0.503	0.435

	sharp	SI	NIQE	SSEQ
Encryption	0.452	0.453	0.473	0.457
Extraction	0.432	0.428	0.537	0.754

This is already a horrible start. Only BLIINDS-II and SSEQ can even order two images based on quality.

Evaluation—Confidence: Monotonicity I

IVC SelectEncrypt database

	VIF	BIQAA	BLIINDS-II	BRISQUE	GPC
iwind ec	0.518	0.194	0.584	0.598	0.098
iwind nec	0.732	0.687	0.717	0.143	0.615
resolution	0.823	0.393	0.286	0.107	0.571
trad	0.913	0.805	0.805	0.560	0.676
truncation	0.832	0.407	0.685	0.885	0.868

	sharp	SI	NIQE	SSEQ
iwind ec	0.001	0.001	0.012	0.699
iwind nec	0.516	0.516	0.676	0.648
resolution	0.429	0.607	0.036	0.107
trad	0.876	0.907	0.764	0.437
truncation	0.868	0.868	0.797	0.558

Evaluation—Confidence: Monotonicity II

We also evaluated on the low quality image of the LIVE database (not shown, see paper).

- The results were slightly better (for all QI)
- Overall the no reference quality indices still hat a lot of problems.

The difference between the referenced quality index and the non-referenced quality indices is quite stark here.

- Each of the no-reference indices has at least one testset for which it fails.

Evaluation—Confidence: Correspondence I

IVC SelectEncrypt Database

	VIF	BIQAA	BLIINDS-II	BRISQUE	GPC
$\mu(\mathcal{C}_D)$	0.277	0.322	0.569	0.597	0.668
$\sigma(\mathcal{C}_D)$	0.098	0.239	0.201	0.190	0.306
Shape	Bias Hi	Stable	Bias Hi	Bias Hi	Stable
Shape †	Bias Lo	Stable	Bias Hi	Bias Hi	Bias Lo

	sharp	SI	NIQE	SSEQ
$\mu(\mathcal{C}_D)$	0.699	0.699	0.415	0.554
$\sigma(\mathcal{C}_D)$	0.333	0.334	0.166	0.130
Shape	Stable	Stable	Bias Lo	Bias Hi
Shape †	Bias Lo	Bias Lo	Bias Lo	Bias Hi

†…LIVE Database

Evaluation—Confidence: Correspondence II

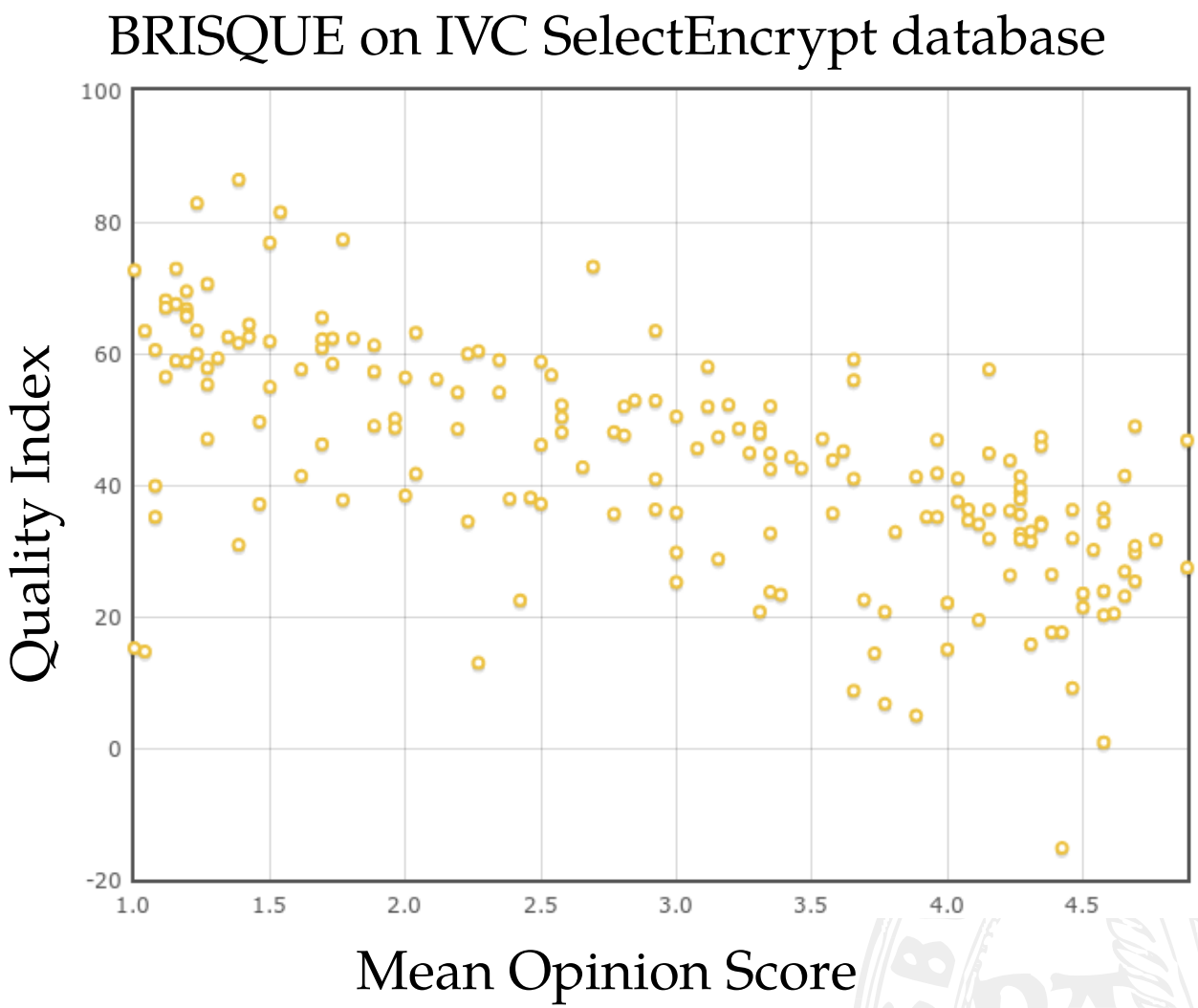
We also evaluated on the low quality image of the LIVE database (not shown, see paper).

- The results are similar for almost all cases.
- The main difference is that the signal shape differs between the databases.
  If that happens, the shape overall has to be counted as unstable.

Difference between full- and no-reference QI:

- Very high values for μ for the NR QIs.
- Overall a more in agreement between databases (both LEG and VIF have to be counted as unstable).

Evaluation—An example



$$\mu(\mathcal{C}_D) = 0.597, \sigma(\mathcal{C}_D) = 0.190,$$
$$SROC_{LOWQuality} = 0.2246, SROC_{FULL_Quality} = 0.6692$$

Learning Low Quality

- No-reference QI are worse than full-reference QI.
- But NR-QI can learn!

What do we learn?

- Learn on the IVC-SelectEncrypt database
- Using BRISQUE because research friendly (open source).
- Using cross validation (one image evaluation, rest training)
- Two modes fitness functions:
  - *BRISQUE low* uses low quality SROC
  - *BRISQUE cross* uses full quality SROC

Learning Low Quality—Results

SROC on	BRISQUE	BRISQUE cross	BRISQUE low
iwind ec	0.598	0.485	0.408
iwind nec	0.143	0.709	0.676
resolution	0.107	0.321	0.393
trad	0.560	0.437	0.723
truncation	0.885	0.657	0.750
full-qualtiy	0.642	0.767	0.745
low-qualtiy	0.304	0.364	0.636

- Overall performance improves, but there is a tradeoff:
  - test sets with bad performances improve (iwind nec and resolution)
  - at the cost well performing test sets (truncation)
- This shifts all test sets towards a mediocre score.
- **Recommendation**: train as specifically as possible.

Conclusion

How do NR-QIs as quality sensitivity index?

- The no-reference VQIs behave overall very similar to most FR-VQIs.
- VIF and LEG outperform the NR-VQIs.
- Recommended to use VIF or, if time is a constraint, LEG.

Training NR-QIs—Yay or Nay?

- Learning improves the overall performance quite a lot.
- If the specific application is known and training data is available the trained NR-VQI can be a better choice.
- As a general purpose VQI for security metrics the VIF and LEG still are a better choice.