On Confidence and Response Times of Human Observers in Subjective Image Quality Assessment

Ulrich Engelke[†], Anthony Maeder[*], Hans-Jürgen Zepernick[†]

[†] Blekinge Institute of Technology (BTH)
Radio Communications Group

[*] University of Western Sydney (UWS)
School of Computing and Mathematics

ICME'09, New York City, 1st July 2009
(Also presented at VQEG Meeting, Berlin, 24th June 2009)

Outline

Outline

Motivation

- Mean Opinion Scores (MOS) considered as most reliable measures of perceived visual quality.
- MOS are widely used to design objective visual quality metrics.
- Rating quality is not necessarily an easy task, in particular, when a variety of artifacts is present.
- Confidence intervals provide additional information regarding the agreement between observers.
- Disagreement can be due to:
  i. Visual quality is hard to judge (for instance very local artifacts).
  ii. Detection and preference of artifacts differs between observers (for instance, some prefer blur others blocking).
- In this respect, additional information regarding observer confidence is of interest.

Confidence and Response Times

Confidence Score (CS)

- Score quantifying how confident an observer was when giving a particular quality score (QS)
- Provided by the observer
- Direct measure of confidence
- May be inconvenient in some cases

Response Time (RT)

- Time required by the observer to give a particular QS
- Measured by the experimenter
- Indirect measure of confidence
- Non-intrusive to the observer

Aims

- Establish a relationship between QS, CS, and RT.
- Model prediction of mean confidence scores (MCS).

Confidence and Response Times

Confidence Score (CS)

- Score quantifying how confident an observer was when giving a particular quality score (QS)
- Provided by the observer
- Direct measure of confidence
- May be inconvenient in some cases

Response Time (RT)

- Time required by the observer to give a particular QS
- Measured by the experimenter
- Indirect measure of confidence
- Non-intrusive to the observer

Aims

- Establish a relationship between QS, CS, and RT.
- Model prediction of mean confidence scores (MCS).

Confidence and Response Times

Confidence Score (CS)

- Score quantifying how confident an observer was when giving a particular quality score (QS)
- Provided by the observer
- Direct measure of confidence
- May be inconvenient in some cases

Response Time (RT)

- Time required by the observer to give a particular QS
- Measured by the experimenter
- Indirect measure of confidence
- Non-intrusive to the observer

Aims

- Establish a relationship between QS, CS, and RT.
- Model prediction of mean confidence scores (MCS).

Hypotheses

H1 It is easier to rate an image if its quality is either very good or very bad while images of medium quality are harder to judge. As a measure of difficulty when judging image quality we consider a confidence score given by a human observer.

H2 The confidence of a human observer when rating the quality of an image is strongly related to the response time of the quality rating. As such, we expect a longer response time for images that are harder to judge.

H3 Observer confidence can be predicted with reasonable accuracy based on the given quality score in combination with the response time measured. Such a confidence prediction may be used as a measure of reliability of a particular MOS.

Hypotheses

H1 It is easier to rate an image if its quality is either very good or very bad while images of medium quality are harder to judge. As a measure of difficulty when judging image quality we consider a confidence score given by a human observer.

H2 The confidence of a human observer when rating the quality of an image is strongly related to the response time of the quality rating. As such, we expect a longer response time for images that are harder to judge.

H3 Observer confidence can be predicted with reasonable accuracy based on the given quality score in combination with the response time measured. Such a confidence prediction may be used as a measure of reliability of a particular MOS.

Hypotheses

H1 It is easier to rate an image if its quality is either very good or very bad while images of medium quality are harder to judge. As a measure of difficulty when judging image quality we consider a confidence score given by a human observer.

H2 The confidence of a human observer when rating the quality of an image is strongly related to the response time of the quality rating. As such, we expect a longer response time for images that are harder to judge.

H3 Observer confidence can be predicted with reasonable accuracy based on the given quality score in combination with the response time measured. Such a confidence prediction may be used as a measure of reliability of a particular MOS.

Outline

Introduction

Subjective Quality Experiment

Quality scores, confidence scores, and response times

Prediction of Observer Confidence

Conclusions

Test Material

Images

- 7 reference images.
- Simulation model to create test images.

Simulation model

- JPEG source encoder.
- (31,21)BCH channel encoder.
- BPSK modulator.
- Rayleigh flat fading channel with AWGN.
- $E_b/N_0 = 5$dB.

Reference Images

Test Material

Images

- 7 reference images.
- Simulation model to create test images.

Simulation model

- JPEG source encoder.
- (31,21)BCH channel encoder.
- BPSK modulator.
- Rayleigh flat fading channel with AWGN.
- $E_b/N_0 = 5$dB.

Blocking

Test Material

Images

- 7 reference images.
- Simulation model to create test images.

Simulation model

- JPEG source encoder.
- (31,21)BCH channel encoder.
- BPSK modulator.
- Rayleigh flat fading channel with AWGN.
- $E_b/N_0 = 5$dB.

Blur

Test Material

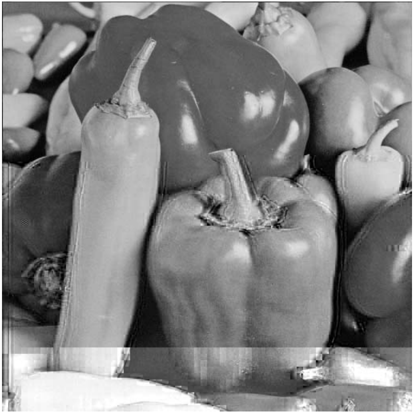
Images

- 7 reference images.
- Simulation model to create test images.

Simulation model

- JPEG source encoder.
- (31,21)BCH channel encoder.
- BPSK modulator.
- Rayleigh flat fading channel with AWGN.
- $E_b/N_0 = 5\text{dB}$.

Ringing / Intensity Masking



Experiment Procedures

University of Western Sydney
Bringing knowledge to life

- Conducted at University of Western Sydney, Australia.
- Number of participants: 15
- Two sessions of about 10 minutes each.
- Test stimuli: 40 test images + 7 reference images in each session.
- Presentation time: 8s/image, 5s/grey screen.

Rating Scales

- During grey screen participants were asked to give a quality score (QS) and confidence score (CS).
- Here, CS quantifies the degree of difficulty to provide the corresponding QS.
- Rating scales for QS and CS:

QUALITY SCORE

Very Good	5
Good	4
Fair	3
Bad	2
Very Bad	1

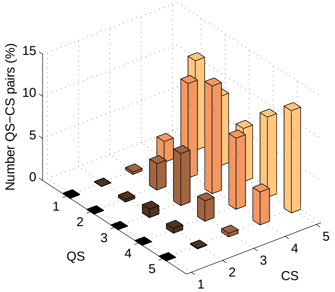
CONFIDENCE SCORE

Very High	5
High	4
Medium	3
Low	2
Very Low	1

- Experimenter measured response time (RT) needed to give both QS and CS.

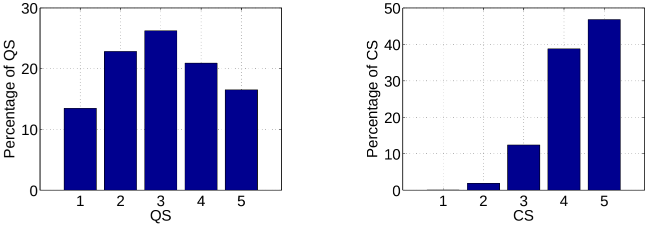
Outline

Occurrence of pairs of QS and CS



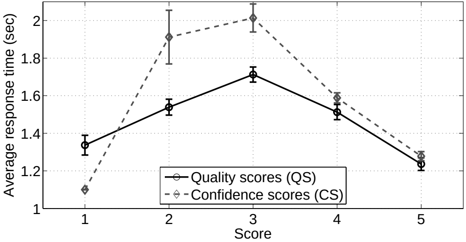
- High confidence (CS=5) at either end of the quality scale (QS=1/QS=5). High confidence ratings drop towards the middle of the quality scale.
- Lower confidence (CS$\leq$ 4) is predominant in the middle of the quality scale.
- These observations ratify hypothesis H1: images of medium quality are harder to judge.

Overall Percentage of QS and CS



- Whole spectrum of QS is covered.
- Strong tendency towards high CS scores.

Average RT for QS and CS



- Average RT increases with decreasing CS (CS=1 may constitute an outlier).
- Average RT increases towards the middle of the quality scale. This is in agreement with decreasing CS towards the middle of QS.
- These observations ratify hypothesis H2: confidence is related to response time.
- As such, RT may contribute information about observer confidence.

Definitions

- The above findings indicate strong relationships between QS, CS, and RT.
- Consider mean scores over all participants as follows
  - Mean quality (opinion) score (MOS): μ_{QS}
  - Mean confidence score (MCS): μ_{CS}
  - Mean response time (MRT): μ_{RT}
- CS and RT are related to the distance of QS to the middle of the quality scale $m_{QS} = 3$. We define delta-QS (DQS) as follows

$$\mu_{QS}^{\Delta} = |\mu_{QS} - m_{QS}| \tag{1}$$

Correlations between μ_{QS}, μ_{CS}, and μ_{RT}

- Pearson linear correlation coefficient

$$\rho_P(u,v) = \frac{\sum_{k=1}^{K}(u_k-\bar{u})(v_k-\bar{v})}{\sqrt{\sum_{k=1}^{K}(u_k-\bar{u})^2}\sqrt{\sum_{k=1}^{K}(v_k-\bar{v})^2}} \quad (2)$$

where u_k and v_k represent any combination of μ_{QS}^{Δ}, μ_{CS}, and μ_{RT}.

- Interdependencies in terms of correlations

$$\begin{aligned} \rho_P(\mu_{QS}^{\Delta}, \mu_{CS}) &= 0.825 \\ \rho_P(\mu_{QS}^{\Delta}, \mu_{RT}) &= -0.714 \\ \rho_P(\mu_{CS}, \mu_{RT}) &= -0.696 \end{aligned} \quad (3)$$

Outline

Introduction

Subjective Quality Experiment

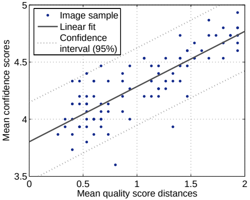
Quality scores, confidence scores, and response times

Prediction of Observer Confidence

Conclusions

Prediction of MCS from DQS or MRT

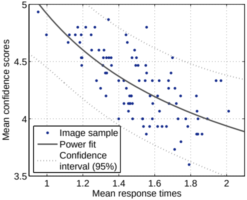
DQS



- Linear function:

$$\mu_{CS}^{(QS)}(a,b) = a + b \cdot \mu_{QS}^{\Delta} \qquad (4)$$

MRT



- Power function:

$$\mu_{CS}^{(RT)}(a,b,c) = a + b \cdot \mu_{RT}^{c} \qquad (5)$$

Table: Prediction function parameters.

	a	b	c
Linear fit (4)	3.802	0.483	-
Power fit (5)	2.679	2.236	-0.829

Combinatorial Prediction Model

- Combinatorial model using weighted L_p-norm ($\rho_P(\mu_{CS}, \mu_{CS}^{pred}) = 0.843$):

$$\mu_{CS}^{pred}(\omega, p) = \left[\omega \cdot (\mu_{CS}^{(QS)})^p + (1-\omega) \cdot (\mu_{CS}^{(RT)})^p\right]^{\frac{1}{p}} \qquad (6)$$

with Minkowski parameter $p \in \mathbb{Z}^+$ and relevance weight $\omega \in [0, 1]$.

- Simple model ($\rho_P(\mu_{CS}, \mu_{CS}^{pred}) = 0.845$):

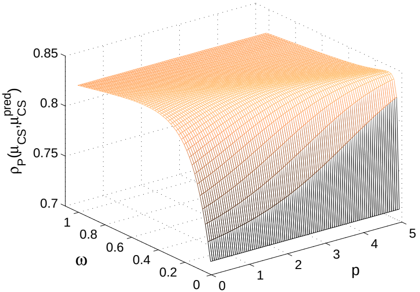
$$\mu_{CS}^{pred}(\omega, p) = \left[\omega \cdot \mu_{QS}^p + (1-\omega) \cdot \left(\frac{1}{\mu_{RT}}\right)^p\right]^{\frac{1}{p}} \qquad (7)$$

- Optimal parameters through exhaustive search:

$$p_{Opt} = 3.036, \qquad \omega_{Opt} = 0.184 \qquad (8)$$

- The correlations ratify hypothesis H3: observer confidence can be predicted with reasonable accuracy based on QS and RT.

Dependence of Model on Parameters p and ω



- The model is strongly dependent on the relevance weight ω and less dependent on the Minkowski parameter p.

Outline

Introduction

Subjective Quality Experiment

Quality scores, confidence scores, and response times

Prediction of Observer Confidence

Conclusions

Conclusions

- We analysed the relationship between QS, CS, and RT as obtained in a subjective experiment.
- We revealed that valuable information about an observers confidence can be derived from both QS and RT.
- We proposed a model for confidence prediction based on QS and RT.
- Future work: analyse relationship of our prediction model to confidence intervals.

Thank you for your attention.

Ulrich Engelke
Blekinge Institute of Technology
Mobile: +46-768-845877
ulrich.engelke@bth.se