Methodology for Objective Metrics Performance Evaluation... 

... and its use for large scale training 

Lukáš Krasula 
lukas.krasula@univ-nantes.fr
Objective Metrics Performance Evaluation

- Comparing subjective vs. automatically predicted scores ($S$ vs. $OM$)
Objective Metrics Performance Evaluation

- Comparing subjective vs. automatically predicted scores \((S \text{ vs. } OM)\)
- Typical measures [ITU-T Rec. P.1401]
  - Pearson Correlation Coefficient
  - Root Mean Squared Error
  - Outlier Ratio
Objective Metrics Performance Evaluation

- Comparing subjective vs. automatically predicted scores ($S$ vs. $OM$)

- Typical measures [ITU-T Rec. P.1401]
  - Pearson Correlation Coefficient
  - Root Mean Squared Error
  - Outlier Ratio

Necessity of mapping to the common scale
Danger of Mapping

- Mapping is not standardized (only required to be monotonic)
- Problems:
Danger of Mapping

- Mapping is not standardized (only required to be monotonic)
- Problems:
  - Different papers provide different results obtained for the same datasets
    - Reproducibility is questionable
Danger of Mapping

- Mapping is not standardized (only required to be monotonic)
- Problems:
  - Different papers provide different results obtained for the same datasets
    - Reproducibility is questionable
  - Mapping can bias the results

<table>
<thead>
<tr>
<th>Correlation for CSIQ database after 3rd order polynomial mapping</th>
<th>SSIM</th>
<th>MS-SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fitting function coefficients optimized with PLCC (VQEG)</td>
<td>0.8575</td>
<td>0.8562</td>
</tr>
<tr>
<td>Fitting function coefficients optimized with RMSE (ITU-T Rec. J.149)</td>
<td>0.8581</td>
<td>0.8859</td>
</tr>
</tbody>
</table>
Rank Order Correlation

- Using Rank Order Correlation Coefficients (Spearman’s and/or Kendall’s)
  - Typical solution to the mapping problem - independency towards the monotonic mapping
Rank Order Correlation

- Using Rank Order Correlation Coefficients (Spearman’s and/or Kendall’s)
  - Typical solution to the mapping problem - independency towards the monotonic mapping

- However...
  - Considering subjective data to be deterministic
Rank Order Correlation

● Using Rank Order Correlation Coefficients (Spearman’s and/or Kendall’s)
  ○ Typical solution to the mapping problem - independency towards the monotonic mapping

● However...
  ○ Considering subjective data to be deterministic

![Graph showing rank order correlation]
Rank Order Correlation

- Using Rank Order Correlation Coefficients (Spearman’s and/or Kendall’s)
  - Typical solution to the mapping problem - independency towards the monotonic mapping

- However...
  - Considering subjective data to be deterministic
Rank Order Correlation

● Using Rank Order Correlation Coefficients (Spearman’s and/or Kendall’s)
  ○ Typical solution to the mapping problem - independency towards the monotonic mapping

● However...
  ○ Considering subjective data to be deterministic

![Graph showing stimulus numbers vs scores](image)

What is the correct order?
Novel performance evaluation methodology

- **Goals:**
  - No mapping during the process
  - Considering the uncertainty of the ground truth
Novel performance evaluation methodology

● Goals:
  ○ No mapping during the process
  ○ Considering the uncertainty of the ground truth

● Basic premise:
  ○ Regardless the subjective procedure, we are always able to determine:
Novel performance evaluation methodology

- **Goals:**
  - No mapping during the process
  - Considering the uncertainty of the ground truth

- **Basic premise:**
  - Regardless the subjective procedure, we are always able to determine:

  \[(a) \text{ Are any two stimuli statistically significantly different in quality?} \]

\[
[i,j] \in N \iff \Pr\{ S(i) \neq S(j) \} < 1-\alpha
\]
\[
[i,j] \in D \iff \Pr\{ S(i) \neq S(j) \} \geq 1-\alpha
\]
Novel performance evaluation methodology

● Goals:
  ○ No mapping during the process
  ○ Considering the uncertainty of the ground truth

● Basic premise:
  ○ Regardless the subjective procedure, we are always able to determine:

(a) Are any two stimuli statistically significantly different in quality?

\[
\begin{align*}
[i,j] \in N & \iff \Pr\{ S(i) \neq S(j) \} < 1-\alpha \\
[i,j] \in D & \iff \Pr\{ S(i) \neq S(j) \} \geq 1-\alpha
\end{align*}
\]

(b) If they are, which of them is qualitatively better?

\[
\begin{align*}
[i,j] \in B & \iff \Delta S(i,j) = S(i) - S(j) \geq 0, \ \forall \ [i,j] \in D \\
[i,j] \in W & \iff \Delta S(i,j) = S(i) - S(j) \leq 0, \ \forall \ [i,j] \in D
\end{align*}
\]
Novel performance evaluation methodology:
Proposed Assumptions

- Reliable metric then

  1. Provides **close** scores for **similar** pairs and **distant** scores for **different**

     \[
     |\Delta OM(i,j)| = |OM(i) - OM(j)| \to 0, \quad \forall \ [i,j] \in N
     \]

     \[
     |\Delta OM(i,j)| = |OM(i) - OM(j)| \gg 0, \quad \forall \ [i,j] \in D
     \]
Novel performance evaluation methodology:
Proposed Assumptions

- Reliable metric then
  
  I. Provides **close** scores for **similar** pairs and **distant** scores for **different**

  \[
  |\Delta OM(i,j)| = |OM(i) - OM(j)| \to 0, \quad \forall \ [i,j] \in \mathbb{N}
  \]

  \[
  |\Delta OM(i,j)| = |OM(i) - OM(j)| \gg 0, \quad \forall \ [i,j] \in \mathbb{D}
  \]

  II. Provides **higher** score for qualitatively **better** stimulus

  \[
  \text{sign} \{ \Delta OM(i,j) \} = \text{sign} \{ \Delta S(i,j) \}, \quad \forall \ [i,j] \in \mathbb{D}
  \]
Novel performance evaluation methodology:

Description

$S, CI, OM$

Dataset(s)
Novel performance evaluation methodology:

Description

Dataset(s) \( S, CI, OM \) → Preprocessing

- **Pairs without significant difference in votes**
  \( [i,j] \in N \iff \Pr \{ S(i) \neq S(j) \} < 1-\alpha \)

- **Significantly different pairs**
  \( [i,j] \in D \iff \Pr \{ S(i) \neq S(j) \} \geq 1-\alpha \)
Novel performance evaluation methodology:
Description

\[ [i,j] \in N \iff \Pr\{ S(i) \neq S(j) \} < 1-\alpha \]

\[ [i,j] \in D \iff \Pr\{ S(i) \neq S(j) \} \geq 1-\alpha \]

\[ [i,j] \in W \iff \Delta S(i,j) = S(i) - S(j) \leq 0 \]

\[ [i,j] \in B \iff \Delta S(i,j) = S(i) - S(j) \geq 0 \]
Novel performance evaluation methodology:

Description

Dataset(s)

\[ |ΔOM(i,j)| = |OM(i) - OM(j)| \]

Pairs without significant difference in votes

|ΔOM(i,j)| = |OM(i) - OM(j)|

S, CI, OM

Preprocessing

N

D

Significantly different pairs

W

B

Pairs with negative score difference

Pairs with positive score difference
Novel performance evaluation methodology:

Description

\[ \text{OM}(i) \] - objective metric's score for stimulus \( i \)

\[ \Delta \text{OM}(i,j) = \text{OM}(i) - \text{OM}(j) \]

difference of objective scores for stimuli \( i \) and \( j \)

\[ |\Delta \text{OM}(i,j)| = |\text{OM}(i) - \text{OM}(j)| \rightarrow 0, \ \forall [i,j] \in N \]

\[ |\Delta \text{OM}(i,j)| = |\text{OM}(i) - \text{OM}(j)| \gg 0, \ \forall [i,j] \in D \]
Novel performance evaluation methodology:

Description

- **OM** (i) - objective metric's score for stimulus i
- **ΔOM**(i,j) = OM(i) - OM(j) - difference of objective scores for stimuli i and j

**Dataset(s)**

- S, CI, OM

**Preprocessing**

- N

**Significantly different pairs**

- |ΔOM(i,j)|

**Pairs without significant difference in votes**

- |ΔOM(i,j)|

**Pairs with negative score difference**

- W

**Pairs with positive score difference**

- B

**Outcomes:**

- AUC value showing how well can the criterion distinguish between significantly different and similar stimuli
- Threshold for the criterion's scores difference providing 95% probability that the images are significantly different (i.e., 0.95 percentile of the distribution for similar pairs)
Novel performance evaluation methodology:

**Description**

- **OM** \( (i) \) - objective metric's score for stimulus \( i \)
- \( \Delta OM(i,j) = OM(i) - OM(j) \) - difference of objective scores for stimuli \( i \) and \( j \)

**Outcomes:**
- AUC value showing how well can the criterion distinguish between significantly different and similar stimuli
- Threshold for the criterion's scores difference providing 95% probability that the images are significantly different (i.e. 0.95 percentile of the distribution for similar pairs)

\[
\text{sign} \{ \Delta OM(i,j) \} = \text{sign} \{ \Delta S(i,j) \}
\]
Novel performance evaluation methodology:

Description

- **OM** \((i)\) - objective metric's score for stimulus \(i\)
- **ΔOM** \((i,j)\) = OM\((i)\) - OM\((j)\) - difference of objective scores for stimuli \(i\) and \(j\)

**Outcomes:**

1. **Different vs. Similar Analysis**
   - AUC value showing how well can the criterion distinguish between significantly different and similar stimuli
   - Threshold for the criterion's scores difference providing 95% probability that the images are significantly different (i.e., 0.95 percentile of the distribution for similar pairs)

2. **Better vs. Worse Analysis**
   - Percentage of correct recognition of the qualitatively better stimulus from the pair
   - AUC value showing how well can the criterion recognize qualitatively better stimulus from the pair
Novel performance evaluation methodology:
Advantages

- Goals have been fulfilled
  - There is no mapping involved
  - The uncertainty of the subjective scores is considered
Novel performance evaluation methodology: Advantages

- **Goals have been fulfilled**
  - There is no mapping involved
  - The uncertainty of the subjective scores is considered

- **Moreover...**
  - Universality towards the subjective procedure, scale, and format of the ground-truth data
  - Allows for simple numerical comparisons and testing of statistical significance
  - High statistical power (due to the pair-wise approach)
  - Enables simple and meaningful combination of the data coming from multiple datasets
Novel performance evaluation methodology: Advantages

- **Goals have been fulfilled**
  - There is no mapping involved
  - The uncertainty of the subjective scores is considered

- **Moreover...**
  - Universality towards the subjective procedure, scale, and format of the ground-truth data
  - Allows for simple numerical comparisons and testing of statistical significance
  - High statistical power (due to the pair-wise approach)
  - **Enables simple and meaningful combination of the data coming from multiple datasets**

![Histograms for databases 1, 2, and 1 & 2 showing the distribution of $|\Delta OM|$ values.](image)
Novel performance evaluation methodology: 

Advantages

- **Goals have been fulfilled**
  - There is no mapping involved
  - The uncertainty of the subjective scores is considered

- **Moreover…**
  - Universality towards the subjective procedure, scale, and format of the ground-truth data
  - Allows for simple numerical comparisons and testing of statistical significance
  - High statistical power (due to the pair-wise approach)
  - Enables simple and meaningful combination of the data coming from multiple datasets
Goals have been fulfilled
- There is no mapping involved
- The uncertainty of the subjective scores is considered

Moreover...
- Universality towards the subjective procedure, scale, and format of the ground-truth data
- Allows for simple numerical comparisons and testing of statistical significance
- High statistical power (due to the pair-wise approach)
- Enables simple and meaningful combination of the data coming from multiple datasets
  - No inter-experiment mapping necessary
  - Overall performance can be easily determined
  - Increase of number of training/testing points in orders of magnitude - deep learning etc.
Using the framework for objective metrics training

- Input Features
- Features Combination
- Performance Evaluation
- Output Evaluation
- Resulting weights

Input Datasets

Setting of weights
Using the framework for objective metrics training

Input Features → Features Combination → Performance Evaluation → Output Evaluation → Resulting weights

Setting of weights by numerical optimization

Our framework
Preliminary results

● Publicly available VMAF (Video Multi-Method Assessment Fusion) package
  ○ VMAF features (VIF on 4 scales, Detail Loss, Motion)
● 18 datasets (9 used for training, 9 for testing)
● 1 hidden layer, 6 neurons, RELU activation function
Preliminary results

- Publicly available VMAF (Video Multi-Method Assessment Fusion) package
  - VMAF features (VIF on 4 scales, Detail Loss, Motion)
- 18 datasets (9 used for training, 9 for testing)
- 1 hidden layer, 6 neurons, RELU activation function
Preliminary results

- Publicly available VMAF (Video Multi-Method Assessment Fusion) package
  - VMAF features (VIF on 4 scales, Detail Loss, Motion)
- 18 datasets (9 used for training, 9 for testing)
- 1 hidden layer, 6 neurons, RELU activation function

<table>
<thead>
<tr>
<th>Custom Neural Network:</th>
<th>VMAF (trained on one of the datasets):</th>
</tr>
</thead>
<tbody>
<tr>
<td>--- Test set ---</td>
<td>--- Test set ---</td>
</tr>
<tr>
<td>AUC_DS = 0.7869</td>
<td>AUC_DS = 0.7586</td>
</tr>
<tr>
<td>AUC_BW = 0.9550</td>
<td>AUC_BW = 0.9490</td>
</tr>
<tr>
<td>CC_0 = 0.8963</td>
<td>CC_0 = 0.8951</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>--- Test + Train sets ---</th>
<th>--- Test + Train sets ---</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC_DS = 0.7646</td>
<td>AUC_DS = 0.7230</td>
</tr>
<tr>
<td>AUC_BW = 0.9551</td>
<td>AUC_BW = 0.9469</td>
</tr>
<tr>
<td>CC_0 = 0.8957</td>
<td>CC_0 = 0.8954</td>
</tr>
</tbody>
</table>
Thank you for your attention!
ROC Analysis

Dataset(s) → Preprocessing → Pairs without significant difference in votes → Objective evaluation + preprocessing → Significantly different pairs → Different vs. Similar Analysis

|\Delta OM| (-) → P (-)

Threshold

TN - true negative
TP - true positive
FN - false negative
FP - false positive

True positive rate
TPR = TP / (TP + FN)

False positive rate
FPR = FP / (FP + TN)
ROC Analysis

Dataset(s)

Preprocessing

Pairs without significant difference in votes

Significantly different pairs

Objective evaluation + preprocessing

Different vs. Similar Analysis

$P (-)$

$|\Delta OM| (-)$

$TPR (-)$

$FPR (-)$
ROC Analysis

Dataset(s) -> Preprocessing

Pairs without significant difference in votes

Significantly different pairs

Objective evaluation + preprocessing

Different vs. Similar Analysis

$P(\cdot)$ vs. $|\Delta OM| (\cdot)$

$TPR(\cdot)$ vs. $FPR(\cdot)$
ROC Analysis

1. Dataset(s)
2. Preprocessing
3. Pairs without significant difference in votes
4. Significantly different pairs
5. Objective evaluation + preprocessing
6. Different vs. Similar Analysis

- $P(-)$
- $|\Delta OM|(-)$

- TPR $(-)$
- FPR $(-)$
ROC Analysis

Dataset(s) → Preprocessing

Pairs without significant difference in votes

Significantly different pairs → Objective evaluation + preprocessing

Different vs. Similar Analysis

$P(\cdot)$

$|\Delta OM| (\cdot)$

$TPR(\cdot)$

$FPR(\cdot)$
ROC Analysis

- Dataset(s)
- Preprocessing
- Pairs without significant difference in votes
- Significantly different pairs
- Objective evaluation + preprocessing

Different vs. Similar Analysis

- $P(\cdot)$
- $|\Delta O_M| (\cdot)$

- TPR (\cdot)
- FPR (\cdot)
ROC Analysis

Dataset(s)

Preprocessing

Pairs without significant difference in votes

Objective evaluation + preprocessing

Significantly different pairs

Different vs. Similar Analysis

$|\Delta OM|$ (-)

P (-)

0 THR

TPR (-)

0 1

FPR (-) 1
ROC Analysis

- **AUC = 1**
- **AUC = 0.5**
- **AUC = 0.85**