VQEG Meeting 2014, Stockholm, Sweden

Universität Konstanz

No-Reference Video Quality Assessment Based on Artifact Measurement and Statistical Analysis

Kongfeng Zhu[1], Chengqing Li[1], Vijayan Asari[2], and Dietmar Saupe[1]

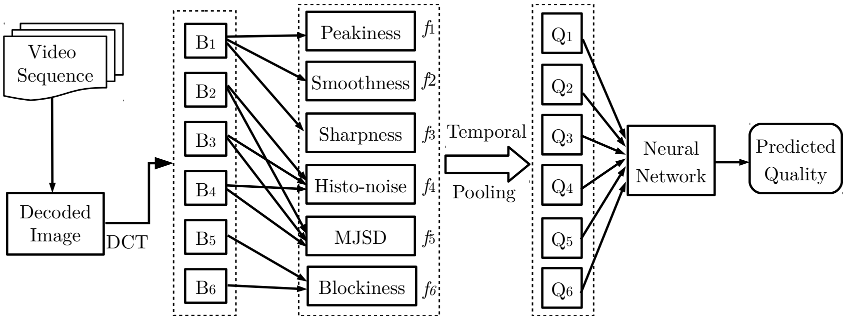
[1]University of Konstanz, Germany

[2]University of Dayton, OH, USA

8th July, 2014

Proposed NR-VQA Model

Universität Konstanz



Step 1

Generate feature maps

Step 2

Extract features

Step 3

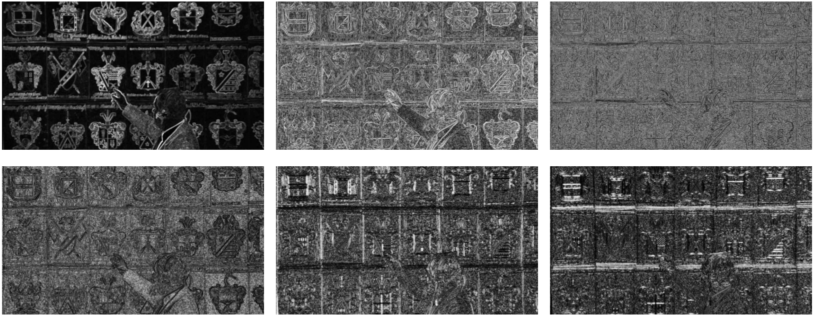
Predict video quality

Universität Konstanz

Step 1: Generate Feature Maps

c_1	c_2	c_3	c_4
c_5	c_6	c_7	c_8
c_9	c_{10}	c_{11}	c_{12}
c_{13}	c_{14}	c_{15}	c_{16}

B_1	B_2	B_3
B_4	B_5	B_6



Step 2: Extract Features

Universität Konstanz

- From B_1:
  1) Kurtosis
  $$f_1(t) = \frac{\sigma_x^4}{E(x-\mu_x)^4} \in (0, 1)$$
  2) Smoothness
  $$f_2(x) = \frac{1}{MN} \left|\{(m,n)|B_1(m,n) < T_L\}\right| \in [0, 1]$$
  3) Sharpness
  $$f_3(x) = \frac{1}{MN} \left|\{(m,n)|B_1(m,n) > T_H\}\right| \in [0, 1]$$
- From B_2, B_3, B_4:
  4) Histogram Noise
  $$f_4(t) = \frac{1}{3} \sum_x [\epsilon_2(x) + \epsilon_3(x) + \epsilon_4(x)] \in [0, 1]$$
  5) Mean Jensen Shannon divergence (MJSD)
  $$f_5(t) = \frac{1}{2} (\mathrm{JSD}(p_2||p_3) + \mathrm{JSD}(p_3||p_4)) \in [0, 1]$$
- From B_5, B_6:
  6) Blockiness
  $$f_6(t) = \frac{B_{\mathrm{MH}} + B_{\mathrm{MV}}}{2} \in [0, 1)$$

Step 3: Predict Video Quality

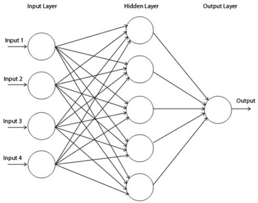
Universität Konstanz

Temporal pooling

- Transform each frame-level feature to a video-level feature: $(f_j(1), f_j(2), \cdots, f_j(t), \cdots, f_j(T_0)) \rightarrow Q_j$
- Minkowski pooling strategy:
  $Q_j = \sqrt[4]{\frac{1}{T_0} \sum_{t=1}^{T_0} f_j(t)^4}$
  where $j = 1, 2, \cdots, 6$, and T_0 is the number of frames

Neural network

- Six inputs: $Q_j, \quad j = 1, 2, \cdots, 6$
- 20 hidden nodes
- One output: predicted MOS



List of databases

Universität Konstanz

Database	a	b	ab	Distortion
IRCCyN video database	60	5	300	H.264/SVC
VQEG HDTV Pool2 database	9	8	72	H.264, MPEG2
LIVE mobile video database	10	4	40	H.264
LIVE video database	10	8	80	H.264, MPEG2

* a stands for the number of references
* b for the number of videos generating from each reference with different quality
* ab for the total number of videos

Validation

Four statistical indices

- LCC: linear correlation coefficient
- SROCC: Spearman's rank ordered correlation coefficient
- RMSE: the root mean squared error
- MAE: the mean absolute error

Cross-validation

- *k*-fold validation for large databases
- leave-*p*-fold-out for small databases

Training models

Universität Konstanz

Linear model (LM)

- $Q = \alpha_0 + \sum_{j=1}^{6} \alpha_j Q_j$
- 7 parameters

Support vector machine (SVM)

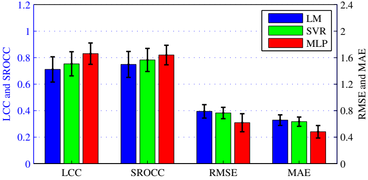
- ε insensitive loss function
- radial basis function kernel

Multilayer perceptron (MLP)

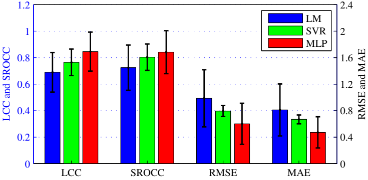
- feed-forward artificial neural network model
- two layers, and 20 nodes in the hidden layer
- the Levenberg-Marquardt backpropagation algorithm

Standard error bar

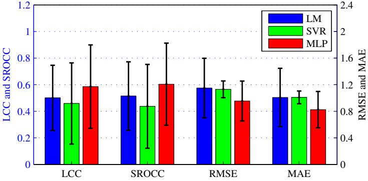
Universität
Konstanz



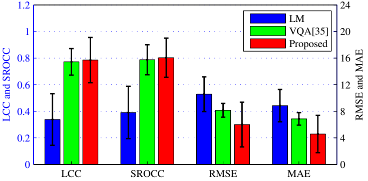
(a) IRCCyN video database



(b) VQEG HDTV Pool2 database



(c) LIVE mobile video database



(d) LIVE video database

Conclusions

Universität Konstanz

Limitations:

- the proposed metric is distortion-specific and data-driven
- prone to over-fitting when the training database is small

Problems:

- Limited number of videos in existing databases
- Cross-database validation is impossible

Solutions ??

- An explicit non-linear mapping with few parameters
- More statistical features?
- Motion analysis?
- Others ?

For more information:

Universität Konstanz

Reference

Kongfeng Zhu, Vijayan K. Asari, Dietmar Saupe, “No-reference quality assessment of H.264/AVC encoded video based on natural scene features”, Mobile Multimedia/Image Processing, Security, and Applications, SPIE Defense, Security, and Sensing, Vol. 8755(4), Baltimore, Maryland, USA, May 2013.

Contact

Kongfeng Zhu: `kongfeng.zhu@uni-konstanz.de`

Visit

`http://www.informatik.uni-konstanz.de/saupe/`

Backup Slides

Step 1: Generate Feature Maps

- Generate DCT map
  A sliding window of size 4×4 moves pixel by pixel
- Define feature maps B_1 to B_6

C_1	C_2	C_3	C_4
C_5	C_6	C_7	C_8
C_9	C_{10}	C_{11}	C_{12}
C_{13}	C_{14}	C_{15}	C_{16}

Feature map	name	description
Unsigned AC	B_1	sum of $C_2, \cdots, C_{16}$
Frequency	B_2	sum of normalized C_2, C_5, C_6
	B_3	sum of normalized $C_3, C_7, C_9, C_{10}, C_{11}$
	B_4	sum of normalized $C_4, C_8, C_{12}, C_{13}, C_{14}, C_{15}, C_{16}$
Orientation	B_5	sum of normalized C_2, C_3, C_4
	B_6	sum of normalized C_5, C_9, C_{13}

Step 1: Generate Feature Maps

Universität Konstanz

An example of the decoded frame

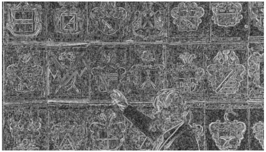
- Only luminance is considered, since the human visual system is more sensitive to luminance than chrominance.
- The size of the decoded frame is $(M+3) \times (N+3)$.
- All feature maps B_1 to B_6 have the same size of $M \times N$.

Step 1: Generate Feature Maps

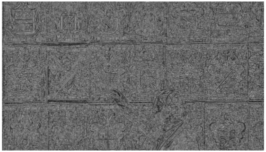
Universität
Konstanz

Examples of feature maps B_1 to B_6

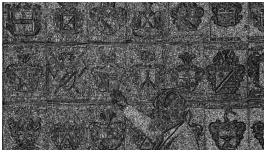
(e) B_1



(f) B_2

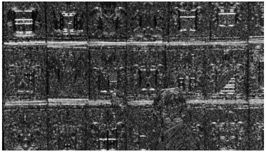


(g) B_3



(h) B_4

(i) B_5

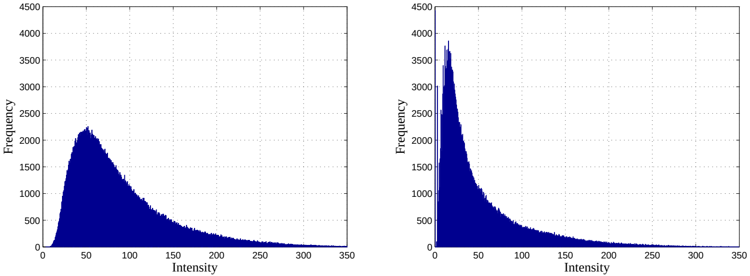


(j) B_6

Step 2: Extract Features from B_1

Universität Konstanz

Histogram of feature map B_1



In contrast to the original frame, the distorted frame has:

- a high histogram peak $\Longrightarrow$ large Kurtosis
- a high frequency around zero $\Longrightarrow$ large smooth area
- low frequency at high intensities $\Longrightarrow$ small sharp area

Step 2: Extract Features from B_1

Universität Konstanz

1) Kurtosis

$$f_1(t) = \text{Kurtosis} = \frac{E(x-\mu_x)^4}{\sigma_x^4} \in [1, \infty)$$

2) Smoothness

$$f_2(x) = \text{Smoothness} = \frac{1}{MN} \left|\{(m,n)|B_1(m,n) < T_L\}\right| \in [0,1]$$

3) Sharpness

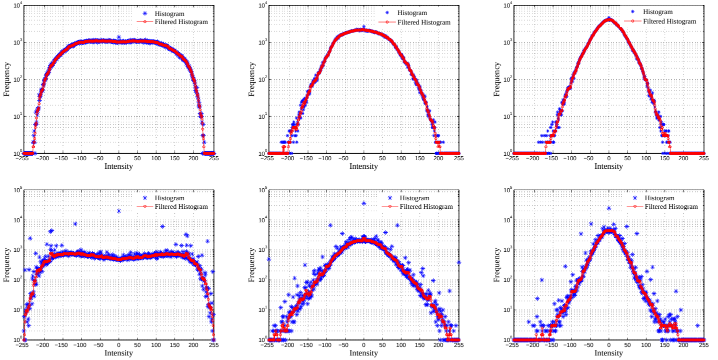
$$f_3(x) = \text{Sharpness} = \frac{1}{MN} \left|\{(m,n)|B_1(m,n) > T_H\}\right| \in [0,1]$$

where

- x is the intensity, μ_x is the mean of x, and σ_x is the standard deviation
- $|A|$ denotes the number of elements of the set A
- $M \times N$ is the size of feature map B_1
- $T_L = 1$ and $T_H = 300$ in the experiment

Step 2: Extract Features from B_2, B_3, B_4

Universität
Konstanz



Histograms on top row are very noisy in comparison to those in bottom, while all the filtered histograms are roughly bilaterally symmetric.

Step 2: Extract Features from B_2, B_3, B_4

Universität Konstanz

4) Histogram noise

$$f_4(t) = \frac{1}{3}\sum_x [\epsilon_2(x) + \epsilon_3(x) + \epsilon_4(x)] \in [0, 1].$$

where

- The histogram noise of band $\mathbf{B}_i$
  $$\epsilon_i(x) = \frac{|\psi_i(x) - \bar{\psi}_i(x)|}{\sum_x \psi_i(x)}, \quad i = 2, 3, 4.$$
- $\psi_i(x)$ for the noisy histogram of band $\mathbf{B}_i$
- $\bar{\psi}_i(x)$ for the filtered version of $\psi_i(x)$

Step 2: Extract Features from B_2, B_3, B_4

Universität Konstanz

It is observed that the *similarity* between two adjacent frequency maps of natural video is *decreased* due to lossy compression.

5) Mean Jensen Shannon divergence (MJSD)

$$f_5(t) = \frac{1}{2}\left(D_{\mathrm{JS}}(p_2||p_3) + D_{\mathrm{JS}}(p_3||p_4)\right) \in [0, 1],$$

where

- $p_2(x), p_3(x)$, and $p_4(x)$ are the smoothed probability density functions of $\mathbf{B}_2$, $\mathbf{B}_3$, and $\mathbf{B}_4$, respectively.
- $D_{\mathrm{JS}}(p||q)$ is the Jensen Shannon divergence, which measures the "distance" between two probability distributions $p(x)$ and $q(x)$.

Step 2: Extract Features from B_5, B_6

Universität Konstanz

6) Blockiness

$$f_6(t) = \frac{1}{2}\left(P_{\mathrm{LKH}} + P_{\mathrm{LKV}}\right) \in (0, 1]$$

- Apply a sum operation along each row in B_6
  $\phi_{\mathrm{H}}(m) = \sum_{n=0}^{N-1} \mathbf{B}_6(m, n), \quad m = 0, ..., M-1$
- Take the 1-D DFT and consider the magnitude
  $$\Phi_{\mathrm{H}}(l) = \left| \sum_{m=0}^{M-1} \phi_{\mathrm{H}}(m) \exp\left(-\frac{j2\pi ml}{L}\right) \right|$$
- The horizontal blockiness measurement (block size $s \times s$)
  $$P_{\mathrm{H}} = \frac{1}{S/2-1} \sum_{s=1}^{S/2-1} \log_{10}\left(\Phi_{\mathrm{H}}\left(\frac{L}{S} \cdot s\right) + 1\right) \in [0, \infty)$$
- $P_{\mathrm{LKH}} = \dfrac{1}{1 + P_{\mathrm{H}}} \in (0, 1]$
- Repeat the process along each column in B_5 for P_{LKV}