Image and Video Processing
Laboratory 5
Subjective and Objective Quality Assessment

Dr. Martin Řeřábek, Evangelos Alexiou, Prof. Touradj Ebrahimi

Reports submission – !!! read carefully !!! Students will have to produce and submit a report two weeks after completion of each laboratory session. An electronic version of the report is submitted via Moodle's interface for uploading assignments. There will be a hard deadline for each submission. Late submissions will have to be justified and explained to responsible assistant by email. A structure of the report is explained below. In addition to the report, the source code produced should also be submitted in electronic form. The report and the source code should be submitted in **ONE ZIP** file using Moodle platform.

Structure of the reports and m-files

- the name of the final ZIP file will be: `lab_`*number-of-lab_your-surname*`.zip`
- the final ZIP file will contain
  - final m-file called `run_lab_`*number-of-lab_your-surname*`.m`
  - all created m-files and functions which are necessary to obtain the results
  - all source images and data needed for successful running of the final m-file
  - an electronic version of your report in **PDF** format
- the structure of the final m-file
  - the final m-file will contain all necessary commands and functions, by running this m-file one must get all results[1]
  - submitted m-files will be commented
  - each figure will be properly titled
  - the final m-file will be divided into cells[2] according to the example bellow
- the parts of program codes should not be included in the final report
- description of the results should be a part of the final report
- don't forget to answer all the questions in your report
- check whether your final m-file can be launched without problems - only then submit your report

[1] of course it is upon you whether you want to include everything in the final m-file, or create different functions which you will call in the final m-file

[2] those who do not know how to use the cells in Matlab to create the program code more transparent and easier to read, they should look at `http://www.mathworks.com/demos/matlab/developing-code-rapidly-with-cells-matlab-video-tutorial.html`

An example of final m-file

```
% Lab (number of lab) - your name and date

%% Exercise (number of exercise) - "Name of exercise"

%your own program code with your coments
a = imread('picture.tif');
%all figures will be properly titled
figure('name','Name Of The Figure')
imshow(a,[]);

%% Exercise (number of exercise) - "Name of exercise"

%your own program code with your coments
[output1, output2, .., outputN] = function_name (input1, input2,..,inputN);
```

0 Introduction

0.1 Subjective quality assessment

Measurement of perceived quality plays a fundamental role in the context of multimedia services and applications. Quality evaluation is needed in order to benchmark image, video, and audio processing systems and algorithms, to test end-devices performance, and to compare and optimize algorithms and their parameters setting. As human subjects usually act as end-users of digital content, subjective tests are performed, where a group of people is asked to rate the quality of the multimedia material, the overall multimedia experience, or the level of impairment by using the same rating scale.

A commonly used scale to evaluate quality is the five-level quality scale:

5 - *Excellent*, 4 - *Good*, 3 - *Fair*, 2 - *Poor*, 1 - *Bad*

A commonly used scale to evaluate level of impairment is the five-level impairment scale:

5 - *Imperceptible*, 4 - *Perceptible, but not annoying*, 3 - *Slightly annoying*, 2 - *Annoying*, 1 - *very annoying*

The limited group of subjects should be a representative sample of the entire population of end-users for the application under analysis. The subjective results are then statistically analyzed in order to understand whether it is possible to draw general conclusions which are valid for the entire population. Subjective quality assessment experiments have to be carried out with scientific rigor in order to provide valid and reliable results.

0.1.1 Test methodology

Several methodologies have been proposed by international standardization bodies for the subjective quality evaluation of moving images, for more details, please refer to [1]. Within this lab session, you will follow the subjective assessment protocol based on ITU-T P.910 for lossy evaluations, namely ACR-HR: Absolute Category Rating with Hidden Reference [2].

0.1.2 Processing of subjective data

In order to interpret the results of subjective quality assessment experiments, the scores assigned by all the observers taking part to a test are averaged in order to obtain the mean opinion score (MOS) for each stimulus:

$$MOS_j = \frac{\sum_{i=1}^{N} m_{ij}}{N} \tag{1}$$

where N is the number of subjects and m_{ij} is the score by subject i for the stimulus j.

Apart from the MOS, another quantity is also computed, which is the confidence interval (CI) of the estimated mean. The CI provides information upon the relationship between the estimated mean values based on a sample of the population (i.e., the limited number of subjects who took part in the experiment) and the true mean values of the entire population. Due to the small number of subjects (usually around 15) the $100 \times (1 - \alpha)\%$ CI is computed using the Student's t-distribution, as follows:

$$CI_j = t(1 - \alpha/2, N) \cdot \frac{\sigma_j}{\sqrt{N}} \tag{2}$$

where $t(1 - \alpha/2, N)$ is the t-value corresponding to a two-tailed Student's t-distribution with $N - 1$ degrees of freedom and a desired significance level α (equal to 1-degree of confidence). N corresponds to the number of subjects, and σ_j is the standard deviation of the scores assigned to the stimulus j. The interpretation of a confidence interval is that if the same test is repeated for a large number of times, using each time a random sample of the population, and a confidence interval is constructed every time, then $100 \times (1 - \alpha)\%$ of these intervals will contain the true value. **Usually, for the analysis of subjective results, the confidence intervals are computed for α equal to 0.05, which corresponds to a degree of confidence of 95%.**

0.2 Objective quality assessment

Although highly informative and reliable, subjective experiments are difficult to design and time-consuming. Furthermore, they cannot be applied when real-time in-service quality evaluation is needed. In order to reduce the effort of subjective testing and overcome its limitations, algorithms, i.e., objective metrics, have been developed in literature to estimate the outcome of the subjective tests. These quality metrics try to automatically and reliably predict the quality of the multimedia content, as perceived by the human end-user.

Objective quality metrics available in literature can be divided in three different categories according to the availability of the original, i.e., reference, signal: full reference (FR) metrics, when both original and processed signals are available; reduced reference (RR) metrics, when besides the processed signal, a description of the original signal and some parameters are available; no-reference (NR) metrics, when only the processed signal is available.

0.2.1 Peak signal-to-noise ratio (PSNR)

Although many metrics have been proposed in literature, the most used algorithm for assessing visual quality of images or video sequences is still the simple pixel based peak signal-to-noise ratio (PSNR). The PSNR is a single frame based metric defined as:

$$PSNR = 10 * log_{10} \frac{(2^B - 1)^2}{MSE} \tag{3}$$

where B is the bit depth of the image component ($B = 8$ for 8 bits per pixel image components) and the mean square error (MSE) is defined as:

$$MSE = \frac{1}{MN} \sum_{y=1}^{M} \sum_{x=1}^{N} (Ref(x, y) - Proc(x, y))^2 \tag{4}$$

being M and N the dimensions of the image component, Ref the original, i.e., reference, image component and $Proc$ the processed image component. The PSNR is a full reference (FR) metric, because both the original, i.e., reference, and processed signals are assumed to be available as input to the metric: the two signals are compared to evaluate the quality of the processed one.

Usually for color video sequences, only the luminance component of the frames is considered, and the video PSNR is computed by averaging the MSE across the luminance components of all the frames. For our evaluation purposes PSNR was computed on both luminance channel $PSNR$ and RGB channels $PSNR_{RGB}$.

0.2.2 Structural similarity index (SSIM)

Another full reference metric which is nowadays widely used is the structural similarity index (SSIM). This metric is built on the hypothesis that the human visual system is adapted to extract structural information from the scene. Structural information can be defined as “the attributes that represent the structure of objects in the scene, independent of the average luminance and contrast”. Thus, the perceived image distortion can be approximated by the structural information change detected between the reference and the processed image. The similarity measure compares the original and the distorted signal considering three main features of images: the luminance, the contrast and the structure. Considering the luminance channel of the reference and test images, these three features are modeled respectively as: mean pixel value, standard deviation of pixel values, and structural map computed by subtracting the mean value from the luminance component and normalizing this difference by the standard deviation. The SSIM map is defined by multiplying the values of three comparison functions: the luminance comparison function, the contrast comparison function and the structural comparison function. Each comparison function is computed as the correlation between the corresponding feature in the reference and the test images. The SSIM metric is applied locally to the image, introducing sliding windows. As result, a SSIM index map is produced, and in order to obtain a representative quality value for the whole image, the mean SSIM value is calculated by computing the mean value over the whole picture.

0.2.3 Multi-scale structural similarity index (MS-SSIM)

The multi-scale structural similarity index (MS-SSIM) is an extension of the SSIM to take into account the fact that the perceivability of image impairments is different depending on the sampling density of the image signal, e.g., as influenced by the viewing distance. To take this into account, the similarity scores are calculated at different spatial scales. The core operation is similar to SSIM: contrast and structural scores are calculated at each scale and the luminance score is calculated at the lowest scale. The factors for combining these scores where found by experiments with human observers.

0.2.4 Visual information fidelity criterion (VIF)

The visual information fidelity criterion (VIF) is another full reference metric which generates objective scores based on a measurement of the mutual information between the test and reference images. VIF uses fundamentally Gaussian models of the wavelet coefficients of the test and reference images that reduce the mutual information measurement to a local signal-to-noise ratio (SNR) in the wavelet domain. VIF uses a multi-scale approach, similarly to MS-SSIM. VIF employs both an energy-based and an edge-based analysis of the reference and test images to estimate the subjective score of the test image. The local SNR calculation at multiple image scales provide an energy-based analysis. The weights applied to the individual image scales adjust the influence of VIF’s edge-based analysis to the overall objective score. Greater emphasis on finer scale analysis increase the role of the edge-based analysis, while greater emphasis on the coarser image scale analysis increases the contribution of the energy-based analysis.

0.2.5 HDR-VDP2.2: High Dynamic Range Visible Difference Predictor [3, 4, 5]

The original HDR-VDP metric [3] was the first metric designed for HDR content. It is an extension of the VDP model [6] that considers a light-adaptive contrast sensitivity function (CSF), which is necessary for HDR content as the ranges of light adaptation can vary substantially. The metric was further extended [4] with different features, including a specific model of the point spread function (PSF) of the optics of the eye, as human optical lens flare can be very strong in high contrast HDR content. The front-end amplitude non-linearity is based on integration of the Weber-Fechner law. HDR-VDP is a calibrated metric and takes into account the angular resolution. The metric uses a multi-scale decomposition. A neural noise block is defined to calculate per-pixel probabilities maps of visibility and the predicted quality metric.

0.2.6 CIEDE2000: Color difference metric

Includes weighting factors for lightness, chroma, and hue (like the CIE1976 L*a*b* perceptual space). Also includes factors to handle the relationship between chroma and hue.

0.2.7 FSIM: Feature Similarity Index

Objective metric based on SSIM adding a comparison of low-level feature sets between the reference and the distorted images. It analyzes the high phase congruency extracting highly informative features and the gradient magnitude, to encode the contrast information. This analysis is complementary and reflects different aspects of the HVS in assessing the local quality of an image. For our purposes FSIM metric was computed for both luminance channel $FSIM$ and chrominance channels $FSIM_C$.

0.3 Comparison of objective and subjective results

The ground truth data gathered through subjective tests is used to benchmark the performance of objective metrics. Usually two attributes are considered in order to compare the prediction performance of the different metrics with respect to subjective ratings:

- *Accuracy* is the ability of a metrics to predict subjective ratings with the minimum average error. It is measured by means of the root-mean-square error (RMSE) and Pearson correlation coefficient (CC), which quantifies the linear correlation between the MOS values and the predicted values. The Pearson correlation coefficient is defined as:

$$CC = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}} \tag{5}$$

  being x_i, y_i, $i = 1, \ldots, n$ the two datasets of n values each, to be compared.

  The value of CC ranges in the $[-1, 1]$ interval, where a value close to 1 (-1) indicates the strongest positive (negative) correlation.

- *Monotonicity* measures if an increase (decrease) in one variable is associated with an increase (decrease) in the other variable, independently of the magnitude of the increase (decrease). It is measured by means of the Spearman correlation coefficient (RC), which quantifies the monotonicity of the mapping, e.g., how well an arbitrary monotonic function could describe the relationship between the MOS values and the predicted values. It is defined as:

$$RC = 1 - \frac{6\sum_{i=1}^{n}(\mathcal{R}(x_i) - (\mathcal{R}(y_i))^2}{n(n^2-1)}, \tag{6}$$

  being x_i, y_i, $i = 1, \ldots, n$ the two datasets of n values each, to be compared, and $\mathcal{R}(\bullet)$ denotes a ranking relation (sorting) applied to the argument.

0.4 Comparison of subjective experiments

Subjective tests are often conducted in a laboratory environment, where the environmental factors, such as viewing conditions, light conditions, display conditions, etc. are controlled as much as possible to ensure the reproducibility and reliability of the results. The modification of environmental conditions can have a severe impact on the results. To determine whether the results obtained from subjective experiments performed in different environmental conditions are similar, the correlation between the sets of results is measured using the Pearson and Spearman correlation coefficients similarly as in Sec. 0.3. As subjective tests are time consuming and expensive, researchers are now starting to use cheaper and faster alternatives such as crowdsourcing. In this case, a pool of workers is hired, for example through Amazon Mechanical Turk, and redirected to a website providing a GUI in order to perform the subjective evaluation remotely.

Table 1: Description of typical quality levels.

Quality	Description of visual content and artifacts corresponding to different quality levels.
Excellent	There are no visible artifacts and distortions (blurriness, blockiness, color artifacts)even after a detailed analysis of the image.
Good	There are minor artifacts, which may be visible after careful analysis but they are not annoying.
Fair	Artifacts are visible immediately or after short time of content analysis and are somewhat annoying.
Poor	Artifacts are immediately visible and are annoying.
Bad	Artifacts are immediately visible and are very annoying making the rendered image useless.

0.4.1 Privacy and data protection

Subjective evaluations produce various private data needed for further analysis (for example data about the sample of participating subjects). Such data are strictly protected, in some countries, such as Switzerland, even by law. Subjective raw scores and individual results are always randomized, so even the investigators doesn't have information connecting individual participants with concrete answers. Prior the subjective tests, a consent form must be distributed to subjects. It contains basic information about the subjective evaluation and list all relevant details (subject's rights, health risks, etc., ...) regarding the tests. By signing such form, participants formally agree to perform the tests. Consent forms are also kept aside, and cannot be mutually linked, to raw test results. An example of such form is attached at the end of this document and will be given to you prior the subjective test and you will be asked to sign it.

1 Participation on the subjective evaluation of image

Without this task you will not be able to finish the lab 5. In this exercise, the task consists in performing a short subjective experiment in semi-controlled environment. This exercise will help us to collect real raw scores from subjective evaluation. The collected data will be analyzed in the following exercises.

1.1 Tasks

1. Log in one of the computer in CO5.
2. **Set up your display resolution to 1920x1200 pixels.**
3. Start chrome browser.
4. Make sure the window of browser is maximized, i.e., the web browser is spanning the whole screen area. Pressing F11 will do it for you.
5. In the browser, go to `http://grebvm2.epfl.ch/Lab5-2016/index.php` and follow the instructions to complete three short sessions of subjective experiments, where you will evaluate the visual quality of several contents compressed with above mentioned codecs. Detailed general description of what is typical Excelent, Good, ... etc. quality is given in Table 1.
   Please note! You should not refrain from using the extreme scores excellent and bad if they match the descriptions above (see Table 1.

Table 2: Material used in the subjective experiment

(a) Proponents/codecs

1	P01	JPEG
2	P02	HEVC
3	P03	JPEG 2000 (PSNR)
4	P04	JPEG 2000 (visual)
5	P05	Daala
6	P06	JPEG XR (444)
7	P07	WebP
8	P08	JPEG (PSNR)
9	P09	JPEG (visual)
10	P10	JPEG XR (420)

(b) Sequences/contents

1	*Bike*
2	*Cafe*
3	*Honolulu_zoo*
4	*p08*
5	*p26*
6	*Woman*

(c) Bitrates

1	*R1*
2	*R2*
3	*R3*
4	*R4*

2 Comparison of different image compression algorithms

The goal of this exercise is to compare the performance of 10 video compression algorithms proposed by various companies/proponents entitled 'P01 - P10' (for codec names see Table 2) on 6 different contents for 4 different bitrates.

2.1 Tasks

1. Download `all.zip` from moodle "Laboratory 5 - Data all". From the zip file, extracts the MATLAB files containing the following mat files

    - `codec_lut.mat`: codec lookup table (240 × 1 vector)
    - `content_lut.mat`: codec lookup table (240 × 1 vector)
    - `bitrate_lut.mat`: bit rate lookup table (240 × 1 vector)
    - `bitrate_values.mat`: bit rate values in [bpp] (240 × 1 vector)
    - `raw_scores.mat`: raw subjective scores, each column correspond to the ratings of one subject (240 × 21 matrix).
    - `metrics.mat`: containing nine variables `<METRIC-NAME>_values`, each having results of one of the objective metric introduced above. (9 variables, each is 240 × 1 vector).

      Note that the raw scores for this subjective experiment are in range 1-5! Do not re-scale them! It corresponds to scores from 21 subjects. These scores were randomly generated for purposes of the lab 5, so you are able to prepare your scripts. After the subjective experiment described in Section 1 is done, you will be able to download the correct raw scores from moodle. Then you will need to produce correct results for the correct subjective data.

    Each row corresponds to a specific condition, i.e., a specific combination of proponent, sequence, and bitrate, which can be determined using the lookup tables and Table 2.

    Example: `codec_lut(i) = 3`, `content_lut(i) = 2`, `bitrates_lut(i) = 4` corresponds to codec *JPEG 2000 (PSNR)* , *Cafe* content, and *R4* bitrate. The corresponding bitrate and objective metric values are given in `bitrate_values(i)` and `<METRIC-NAME>_values(i)`, respectively. The corresponding subjective scores are given in `raw_scores(i,:)` (following MATLAB notation).

2. Write a function to compute the MOS values and corresponding 95% CIs for each test condition (codec, content, bitrate).

3. Plot the MOS values together with CIs for each content (i.e., on the same graph, plot 10 "MOS versus bitrate" curves corresponding to the 10 proponets; plot one graph for each content).

4. Similarly, plot all objective metrics vs bitrate.

5. **Comment upon the obtained results**:
   - Comment upon the obtained results: Which codec performs better in terms of perceived quality?
   - Which codec performs better in terms of objective quality based on each objective metrics?

6. **When the subjective evaluation, in which you have to take part, is done (Wednesday 7.12. in the afternoon), you will be able to find the correct raw scores on moodle. Don't forget to replace the random raw scores by the real data to provide correct results within your report!!!**

   Useful MATLAB commands: `mean`, `std`, `icdf`, `errorbar`

Examples of the graphs required as a results of this exercise can be found in [7].

3 Benchmarking of objective metrics

The goal of this exercise is to benchmark the different objective metrics described in Section 0.2 using the subjective scores as ground truth. For this purpose, you will use again both subjective scores (subjective evaluation in past and current crowdsourcing experiment) and objective measurements, which were computed offline within corresponding subjective evaluation scenario (subjective evaluation in past and current crowdsourcing experiment).

The data is composed as explained in Section 2

3.1 Tasks

1. All necessary data are available by repeating Task 1 in Section 2
2. Plot the MOS values together with CIs for each metric for all sequences(i.e., on the same graph, plot the "MOS vs objective metric" for all sequences; plot one graph for each metric).
3. For each metric, fit the objective measures to the subjective scores, using linear and cubic fitting.
4. BONUS: Use logistic fitting.
5. Plot the fitted curves on the graphs.
6. Measure the Pearson and Spearman correlation coefficients, as well as the root-mean-square error (RMSE), between the fitted scores and the actual subjective scores and report these values in a table.
7. Comment upon the obtained results: which metric performs better in terms of correlation with the subjective results? what is the influence of the fitting process? what is the influence of the content on the performance of objective metric?

   Useful MATLAB commands: `corr`, `polyfit`

Examples of the graphs required as a results of this exercise can be found in [8].

References

[1] ITU-R BT.500-13, "Methodology for the subjective assessment of the quality of television pictures," International Telecommunication Union, January 2012.

[2] ITU-T P.910, "Subjective video quality assessment methods for multimedia applications," International Telecommunication Union, April 2008.

[3] R. Mantiuk, S. J. Daly, K. Myszkowski, and H.-P. Seidel, “Predicting visible differences in high dynamic range images: model and its calibration,” in *Proceedings of SPIE 5666*, ser. Human Vision and Electronic Imaging X, 2005.

[4] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich, “HDR-VDP-2: A Calibrated Visual Metric for Visibility and Quality Predictions in All Luminance Conditions,” *ACM Transactions on Graphics*, vol. 30, no. 4, Jul. 2011.

[5] M. Narwaria, R. Mantiuk, M. Perreira Da Silva, and P. Le Callet, “HDR-VDP-2.2: a calibrated method for objective quality prediction of high-dynamic range and standard images,” *Journal of Electronic Imaging*, vol. 24, no. 1, p. 010501, 2015.

[6] S. J. Daly, “Visible differences predictor: an algorithm for the assessment of image fidelity,” in *Proceedings of SPIE 1666*, ser. Human Vision, Visual Processing, and Digital Display III, 1992.

[7] P. Hanhart, M. Rerabek, P. Korshunov, and T. Ebrahimi, “Subjective evaluation of HEVC intra coding for still image compression,” in *Seventh International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, 2013.

[8] P. Hanhart, P. Korshunov, and T. Ebrahimi, “Benchmarking of quality metrics on ultra-high definition video sequences,” in *18th International Conference on Digital Signal Processing (DSP)*, 2013.

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

INFORMED CONSENT FORM FOR PARTICIPANTS IN RESEARCH STUDIES

Please complete this form after you have read the Information Sheet and received an explanation about the research.

Project: "*Visual Quality Evaluation of Performance of Image Compression Algorithms.* "
Project supervisor: Prof. Dr. Touradj Ebrahimi
Investigator: Dr. Martin Rerabek, Evangelos Alexiou
EPFL STI IEL GR-EB, ELD 225 (Bâtiment ELD), Station 11, CH-1015 , Lausanne
touradj.ebrahimi@epfl.ch

Thank you for your interest in taking part in this research. Before you agree to take part, the person organizing the research will explain the project to you.
If you have any questions arising from the explanation already given to you, please ask the researcher before you participate in the test. You will be given a copy of this Consent Form to keep and refer to at any time.

The questions asked here have the goal to facilitate the processing of the ratings that will be recorded. Your answers will be treated in a strictly confidential manner.
Last name, first name: ....
You are a □ male □ female
Your date of birth (day/month/year) ....

By signing this form, you attest to be at least 18, not to be under the tutelage or legal guardianship, to have well understood the experiment goal and the task that you will be asked to do, and to freely agree to take part in this study.

- The study supervisor and the experimenter have informed me orally about the purposes of the study.
- I have understood the information about the study. I have received sufficient answers to questions about my participation to this study. On written request, I will receive a copy of the present folder (information, consent form and experiment form).
- I voluntary take part in this study. I can quit this study at any moment without explaining my reasons and without any negative consequences.
- I accept that signals from the experiment are recorded and processed in scientific purposes only and by respecting the confidentiality, in accordance with the CH Federal law on data protection ("Loi fédérale sur la protection des données"- RS 235.1). I also accept that scientific publications are realized from the obtained results.
- I have been informed that possible damage to my health, which is directly related to the above study and is demonstrably the fault of EPFL, is covered by the general liability insurance of EPFL (insurance policy no. 30/5.006.824 of the Baloise Insurance). However, beyond the before mentioned, my health insurance and/or accident insurance will apply.
- I had enough time to think before taking my decision.

□ Check here if you accept that your picture could be used for an eventual scientific publication.
Please check: □ I have read the text above and I agree to take part in this experiment.

Subj ect's signature: .... Date: ....