Evaluation of video quality metrics on transmission distortions in H.264 coded video

Iñigo Sedano, Maria Kihl, Kjell Brunnström and Andreas Aurelius

Abstract—The development of high-speed access networks has enabled a variety of video delivery alternatives over the Internet, for example IPTV and peer-to-peer based video services as Voddler. Consequently, the development of real-time video QoE monitoring methods is receiving large attention in the research community. We believe that the good performing objective metrics using reference information could be used to speed up the development process of real-time video QoE monitoring methods. Thus in this paper we study the accuracy of full-reference objective methods for assessing the quality degradation due to the transmission distortions. We evaluated several well-known publicly-available full-reference objective metrics on the freely available EPFL-PoliMI (Ecole Polytechnique Fédérale de Lausanne and Politecnico di Milano) video quality assessment database, which was specifically designed for the evaluation of transmission distortions. The full-reference metrics are usually evaluated using a reference which is uncompressed. Instead, we study the performance of the metrics when the reference videos are lightly compressed to ensure high quality.

Index Terms—Objective evaluation techniques, Performance evaluation, IPTV & Internet TV

I. INTRODUCTION

The development of high-speed access networks has enabled a variety of video delivery alternatives over the Internet, for example IPTV and peer-to-peer based video services as Voddler [1]. A major problem is that the Quality of Experience (QoE) of the video can be severely affected even by a low packet loss rate. Consequently it is important to allocate necessary network resources to minimize the loss of video information. For doing that it is necessary to monitor and estimate the QoE delivered to user. Accordingly, the development of video quality metrics is receiving large attention in the research community.

Currently, the methods using reference information and

...
are thus similar in quality to the uncompressed original. In real network deployments the uncompressed original sequence is usually not available. Therefore we believe there is an interest in evaluating the performance of the metrics when a compressed reference of a bit lower quality is used instead.

The transmission distortions are simulated at different packet loss rates (PLR) (0.1%, 0.4%, 1%, 3%, 5%, 10%) and two different channel realizations are selected for each PLR. The H.264/AVC reference software encoder adopting the High Profile is used both for encoding and decoding the videos. The compressed videos in the absence of packet losses are used as the reference for the computation of the DMOS (Differential Mean Opinion Score) values. Forty naive subjects took part in the subjective tests.

We have evaluated the performance of the following well-known publicly-available video quality algorithms: Peak Signal to Noise Ratio (PSNR), Structural SIMilarity (SSIM) index [11], Multi-scale SSIM (MS-SSIM) [13], Video Quality Metric (VQM) [15], Visual Signal to Noise Ratio (VSNR) [16], MOtion-based Video Integrity Evaluation (MOVIE) [18], Spatial MOVIE [18] and Temporal MOVIE [18].

The performance of the objective models is evaluated using the Spearman Rank Order Correlation Coefficient, the Pearson Linear Correlation Coefficient, the Root Mean Square Error (RMSE) and the Outlier Ratio. A non-linear regression is done using a monotonic cubic polynomial function with four parameters as recommended by the VQEG [7]. The performance of the different metrics is compared by means of a statistical significance analysis based on the Pearson, RMSE and Outlier Ratio coefficients.

II. METHODOLOGY

A. EPFL-PoliMI video database

Three of the reference videos have a frame rate of 25 fps (cropping HD resolution video sequences down to 4CIF (704 × 576 pixels) resolution and downsampling the original content from 50 fps to 25 fps) while the other three have a frame of 30 fps. As it has been already stated, the reference videos are lightly compressed to ensure high quality in the absence of packet losses. Therefore, a fixed Quantization Parameter between 28 and 32 was selected for each sequence.

The sequences were encoded in H.264/AVC [8] High Profile. B-pictures and Context-adaptive binary arithmetic coding (CABAC) were enabled for coding efficiency. Each frame was divided into a fixed number of slices, where each slice consists of a full row of macroblocks.

The error patterns were generated at six different PLRs [0.1%, 0.4%, 1%, 3%, 5%, 10%] with a two state Gilbert’s model with an average burst length of 3 packets. For each PLR and content, two realizations were selected producing a total of 72 distorted sequences.

The subjective evaluation was done using the 5 point ITU continuous scale in the range [0-5] [9]. 21 subjects participated in the evaluation at the PoliMI lab and 19 at the EPFL lab. More details about the subjective evaluation can be found in [4] [5] [6].

B. Processing of the subjective scores

In [5] the Shapiro-Wilk test was used to verify the normality of distributions of scores across subjects and the results of the test showed that the scores distributions are normal or close to normal.

Although the raw subjective scores were already processed in the EPFL-Polimi database, we processed them in a different way.

A T-Student considering the overall mean and standard deviation of the raw MOS individual scores of each lab showed that at 95% confidence level the data from the two labs can be merge. As an additional verification, the DMOS and confidence interval (CI) values (in this case after normalization, screening and re-scaling) were calculated for each content and distortion type and compared between the two labs, confirming that the data from the two labs can be merged.

First of all, we calculated the difference scores by subtracting the scores of the degraded videos to the score of the reference videos. The difference scores for the reference videos are 0 and so are removed. Accordingly, a lower difference score indicates a higher quality.

Afterwards, the Z-scores were computed for each subject separately by means of the Matlab zscore function. The Z-scores transform the original distribution to one in which the mean becomes zero and the standard deviation becomes one. Each subject may have used the rating scale differently and with different offset. Indeed, this normalization procedure reduces the gain and offset between the subjects.

Subsequently, the outliers were detected according to the guidelines described in section 2.3.1 of Annex 2 of [9] and removed.

Next, the Z-scores were re-scaled to the range [0,5]. The Z-scores are assumed to be distributed as a standard Gaussian. Consequently, 99% of the scores will lie in the range [-3,3]. In our study 100% of the scores lied in that range. The re-scaling was done by linearly mapping the range [-3,3] to the range [0,5] using the following formula:

\[ z' = \frac{5(z + 3)}{6} \]

Finally the Difference Mean Opinion Score (DMOS) of each video was computed as the mean of the re-scaled Z-scores from the 36 subjects that remained after rejection. Additionally, the confidence intervals were also computed.

C. Objective assessment algorithms

The performance of the following video quality algorithms was evaluated on the EPFL-PoliMI video quality assessment database:

- Peak Signal to Noise Ratio (PSNR): PSNR is computed using the mean of the MSE vector (contains the Mean Square Error of each frame). The implementation used is based on the "PSNR of YUV videos" program (yuvpsnr.m) by Dima Pröfrock available in the MATLAB Central file repository [10]. Only the luminance values were considered.

---

[References]

• Structural SIMilarity (SSIM): SSIM is computed for each frame. After that an average value is produced. The implementation used is an improved version of the original version [11] in which the appropriate scale is estimated. The implementation, named ssim.m can be downloaded in the author’s implementation home page [12]. Only the luminance values were considered.

• Multi-scale SSIM (MS-SSIM): MS-SSIM [13] is computed for each frame. Afterwards an average value is produced. The implementation used was downloaded from the Laboratory for Image & Video Engineering (LIVE) at the University Of Texas at Austin [14]. Only the luminance values were considered.

• Video Quality Metric (VQM): The software version 2.2 for Linux used was downloaded from the author’s implementation home page [15]. As for the parameters used: parsing type none, spatial, valid, gain and temporal calibration automatic, temporal algorithm sequence, temporal valid uncertainty false, alignment uncertainty 15, calibration frequency 15 and video model general model. The files were converted to the format required by VQM (Big-YUV file format, 4:2:2) using ffmpeg.

• Visual Signal to Noise Ratio (VSNR): VSNR [16] is computed using the total signal and noise values of the sequence. Only the luminance values were considered. We modified the authors’ implementation available at [17] to extract the signal and noise values in order to sum them separately. Only the luminance values were considered.

• MOtion-based Video Integrity Evaluation (MOVIE): MOVIE [18] includes three different versions: the Spatial MOVIE index, the Temporal MOVIE index and the MOVIE index. The MOVIE Index version 1.0 for Linux was used and can be downloaded from [14]. The optional parameters framestart, frameend or frameint were not used. The default values of the metrics were used for all the metrics. No registration problems occur in the dataset.

D. Statistical analysis

In order to test the performance of the objective algorithms we compute the Spearman Rank Order Correlation Coefficient (SROCC), the Pearson correlation coefficient, the Root Mean Square Error (RMSE) and the Outlier Ratio (OR).

The Spearman coefficient assesses how well the relationship between two variables can be described using a monotonic function. The Pearson coefficient measures the linear relationship between a model’s performance and the subjective data. The RMSE provides a measure of the prediction accuracy. Lastly, the consistency attribute of the objective metric is evaluated by the Outlier Ratio.

The Pearson, RMSE and Outlier Ratio are computed after a non-linear regression. The regression is done using a monotonic cubic polynomial function with four parameters. The function is constrained to be monotonic:

\[ DMOSP = a \cdot x^3 + b \cdot x^2 + c \cdot x + d \]

In the above equation the DMOSP is the predicted value. The four parameters are obtained using the MATLAB function “nlinfit”.

The performance of the metrics is compared by means of a statistical significance analysis based on the Pearson, RMSE and Outlier Ratio coefficients [7].

III. RESULTS

We include the scatter plots of the objective metrics scores vs. DMOS for all the videos in the EPFL-PoliMI video quality database. The fitting function is also plotted.

![Fig. 1. Scatter plot PSNR](image1.png)

![Fig. 2. Scatter plot SSIM](image2.png)
Fig. 3. Scatter plot MS-SSIM

Fig. 4. Scatter plot VQM

Fig. 5. Scatter plot VSNR

Fig. 6. Scatter plot MOVIE

Fig. 7. Scatter plot SPATIAL MOVIE

Fig. 8. Scatter plot TEMPORAL MOVIE
In the table I the values of the coefficients corresponding to all the metrics are shown.

<table>
<thead>
<tr>
<th>OBJECTIVE QUALITY ASSESSMENT ALGORITHMS</th>
<th>Pearson</th>
<th>Spearman</th>
<th>RMSE</th>
<th>Outlier Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR</td>
<td>0.9586</td>
<td>0.9614</td>
<td>0.2195</td>
<td>0.6250</td>
</tr>
<tr>
<td>SSIM</td>
<td>0.9595</td>
<td>0.9696</td>
<td>0.2173</td>
<td>0.5972</td>
</tr>
<tr>
<td>MS-SSIM</td>
<td>0.9642</td>
<td>0.9781</td>
<td>0.2046</td>
<td>0.5972</td>
</tr>
<tr>
<td>VSNR</td>
<td>0.9744</td>
<td>0.9735</td>
<td>0.1733</td>
<td>0.4722</td>
</tr>
<tr>
<td>VQM</td>
<td>0.9619</td>
<td>0.9603</td>
<td>0.2109</td>
<td>0.5417</td>
</tr>
<tr>
<td>MOVIE</td>
<td>0.9650</td>
<td>0.9622</td>
<td>0.2023</td>
<td>0.6250</td>
</tr>
</tbody>
</table>

| SPATIAL MOVIE                          | 0.9814  | 0.9787   | 0.1480 | 0.4583        |
| TEMPORAL MOVIE                         | 0.9243  | 0.9142   | 0.2944 | 0.6111        |

IV. DISCUSSION

The statistical significance analysis based on the RMSE shows that at 95% confidence level all the metrics are statistically better than Temporal Movie. On the other hand, Spatial Movie is statistically better than all the other metrics except VSNR. VSNR is statistically better than PSNR and SSIM. The statistical significance analysis based on the Pearson confirms the lower performance of Temporal Movie and the analysis based on the Outlier Ratio does not provide an indication of the differences between the performances of the metrics. The packet loss did not induce registration problems, explaining partly the high correlation values obtained.

V. CONCLUSIONS

The results show that when the lightly compressed sequences without packet losses are taken as the reference instead of the uncompressed videos, the correlation of the selected objective algorithms is high, being the lowest Pearson correlation coefficient after non-linear regression in our study 0.9243. However, we believe the overall correlation would be lower if the models are evaluated over databases containing more sequences, registration problems, different coding parameters (e.g. flexible macroblock ordering) and error concealment strategies. The statistical analysis based on RMSE shows that at 95% confidence level, the Spatial Movie index shows the highest performance and Temporal Movie the lowest among the studied metrics.

REFERENCES