Comparison of video codecs and coded video sequences quality using the latest objective and subjective assessment methods

Marko Hebar, Peter Planinšič
Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko, Smetanova 17, 2000 Maribor, Slovenija
E-pošta: marko.hebar1@uni-mb.si

Abstract. This paper presents the comparisons of the latest subjective video sequence quality assessment methods versus objective quality assessment using H.261, H.263+ and H.264 advanced video coding (AVC). A Simultaneous Double Stimulus for Continuous Evaluation (SDSCE) method was used for subjective quality assessment. The Peak Signal to Noise Ratio (PSNR) method and the-state-of-the-art Structural SIMilarity index method (SSIM) were used for objective quality assessment. The reconstructed video sequences are obtained using H.261, H.263+, H.264 standards encoded to different bit rates ranging from 512 kbit/s through 4096 kbit/s. Results show that the methods SSIM and PSNR are better for high bit rate video assessment and SDSCE for low bit rate video assessment. The SSIM method gives more realistic results than PSNR method. Also compared are H.261, H.263+ and H.264 which is new generation codec and is superior to the older standard codecs.

Keywords: PSNR, quality, SDSCE, SSIM, video

Primerjava videokodekov in kodiranih videosekvenc z novejšimi objektivnimi in subjektivnimi metodami ocenjevanja


Ključne besede: kakovost, PSNR, SDSCE, SSIM, video

Introduction

The use of digital video sequences has increased over recent years. Although there have been great advances in compression and transmission techniques, impairments often occur throughout introduced along the several stages of a communication system [1], [2]. The video coding H.262 standard [3] was developed about 10 years ago primarily as an extension of prior H.261 standard [4]. It is widely used for the transmission over satellite, cable, terrestrial emission and the storage of high-quality video signals. Moreover, other transmission media such as Cable Modem, Digital Subscriber Line (DSL) or Universal Mobile Telecommunications System (UMTS) offer much lower data rates than broadcast channels, and enhanced coding efficiency can enable the transmission of more video channels or higher quality video presentations within existing digital transmission capacities. Video coding for telecommunication applications has evolved through the development of the H.261 [3], H.262 [4], and H.263 [3], video coding standards. Recently the H.264 standard [5] has also begun to emerge in some applications. For quality evaluation, objective video quality assessment methods have been widely used, such as the peak signal-to-noise ratio PSNR [6] and lately structural similarity index SSIM [7]. The subjective and objective assessment of H.261, H.262,
H.263 or H.264 coded video is necessary, because human visual systems cannot provide accurate results for high bit rates videos. New methods were developed for an objective no-reference (NR) metric at that time [10], [11]. Fuzzy image quality measure was proposed in [13].

In this paper, the used video test sequence “Foreman” and a video sequence “Matrix Revolutions” were coded using H.261, H.263+ and H.264 codec, and assessed using objective and subjective methods. Tests were carried out on sequences coded with a constant bit rate, with one pass, at average bit rates form 512 to 4096 kbit/s. The SDSL method was used for subjective assessment. The PSNR and SSIM methods were used for objective assessment. The main goal was to find out at which average bit rate give the objective methods better results than the subjective methods and which encoding method is appropriate obtaining complexity/efficiency compromise. We tested under optimum conditions, as suggested [9].

The paper is organized as follows. Section 2 and 3 briefly introduces the used video encoders and used assessment methods. The experimental results are presented in section 4, section 5 concludes the paper.

2 Video coders

The standard H.261 is a multimedia standard with specifications for the coding, compression, and transmission of audio, video, and data streams in a series of synchronized, multiplexed packets. The focus of the standard was on the storage of multimedia content on a standard Compact Disc Read-Only Memory (CD-ROM), which supported data transfer rates of 1.4 Mbit/s, and a total storage capability of about 600 MB. The picture format that was chosen was the SIF format (352x288 at 25 noninterlaced frames/s or 352x240 pixels at 30 noninterlaced frames/s). The video coding in H.261 is very similar to the video coding of the H.26X series, spatial coding by taking the Discrete Cosine Transform (DCT) of 8x8 pixel blocks, quantizing the DCT coefficients based on perceptual weighting criteria, storing the DCT coefficients for each block in a zigzag scan, and doing a variable run-length coding of the resulting DCT coefficient stream. Temporal coding was achieved by using the ideas of unidirectional and bidirectional motion-compensated prediction, with three types of pictures resulting, I (intra), P (predictive) or B (bidirectional) predictive pictures. High-quality audio coding is also a part of the H.261 standard and includes sampling rates of 32 kHz, 44.1 kHz, and 48 kHz.

The standard H.262, Part 2, is similar to H.261, but also provides support for interlaced video, the format used by broadcast TV systems. H.262 video is not optimized for low-bit rates of less than 1 Mbit/s, but outperforms H.261 at 3 Mbit/s and above. All standards conforming to H.262 video decoders are fully capable of playing back H.261 video streams. The H.262 systems are used in most high definition HDTV transmission systems, with some enhancements. The H.262 audio Part, defined in part 3 of the standard, enhances H.261 audio by allowing the coding of audio programs with more than two channels. Part 3 of the standard allows this to be done in a backwards compatible way, allowing H.261 audio decoders to decode the two main stereo components of the presentation. In part 7 of the H.262 standard, audio can alternatively be coded in a non-backwards-compatible way, which allows encoders to make better use of the available bandwidth. Part 7 is referred to as H.262 advanced audio coding (AAC). The output bit-rate of an H.262 encoder can be constant or variable, with the maximum bit rate determined by the playback media. For example, the Digital Versatile Disc (DVD) movie maximum is 10.4 Mbit/s. H.263 was developed as an evolutionary improvement based on experience from H.261, the previous International Telecommunication Union - Telecommunication (ITU-T) standard for video compression.

The H.264 project intent was to create a standard that would be capable of providing good video quality. Bit rates should be substantially lower, half or less than what previous standards would need, relative to H262, H.263. This had to be done without too much increase in complexity. Otherwise, the designs impractically would be excessively expensive to implement. An additional goal was to do this in such a flexible way as to allow the standard to be applied to a very wide variety of applications. Both low, high bit rates and low, high resolution video should work well on a very wide variety of networks and systems for broadcast, DVD storage, Real-time Transport Protocol/Internet Protocol (RTP/IP) packet networks, and ITU-T multimedia telephony systems. H.264 contains a number of new features that allow for video compress much more effectively than older standards and to provide more flexibility when applied to a wide variety of network environments. Such key features include: Multi-picture motion compensation, variable block-size motion compensation, six-tap filtering, macro block pair structure, quarter-pixel precision for motion compensation, weighted prediction, an in-loop deblocking filter, an exact-match integer 4x4 spatial block transform, secondary Hadamard transform performed on DC coefficients, spatial prediction from the edges of neighboring blocks for intra coding, context-adaptive binary arithmetic coding, context-adaptive variable-length coding, a common simple and highly-structured variable length coding technique for many of the syntax elements not coded by Context-Adaptive Binary Arithmetic (CABAC) or Context-Adaptive Variable (CAVLC), a network abstraction layer, switching slices, flexible macro block ordering, arbitrary slice ordering, data partitioning, redundant
slices, a simple automatic process for preventing the accidental emulation of start codes, supplemental enhancement information, auxiliary pictures, frame numbering and picture order count [5]. These techniques, along with several others, help H.264 to perform significantly better than any prior standard can. New video codecs based on wavelet transform are under extensive research and development [14], [15]. In next chapters, we will describe objective and subjective methods, starting with subjective method.

3 Assessment methods

3.1 Simultaneous double stimulus for continuous evaluation (SDSCE)

The SDSCE method is the standard method for subjective digital coded video assessment [9].

The group of subjects watches two sequences at the same time, one is the reference sequence and the other is the tested sequence. If the format of the sequences is standard image format or smaller, the two sequences can be displayed side by side on the same monitor, otherwise two aligned monitors should be used, as shown in Figure 1.

Subjects are requested to check the differences between the two sequences and to judge the fidelity of the video information by moving the voting device slider. When the fidelity is perfect, the slider should be at the top of the scale range 100 [9]. When the fidelity is nil, the slider should be at the bottom of the scale 0 as suggested in [9].

There are three different evaluation phases.

The training phase is a crucial part of the chosen SDSCE test method, since subjects must understand their task. Written instructions should be provided to be certain that all the subjects receive exactly the same information.

![Figure 1. Example of an assessed display format.](image1)

Slika 1. Primer prikaza slik ocenjevanja.

The instructions should include explanation about what the subjects are going to see, what they have to evaluate and how they express their opinion.

After the instructions, a demonstration session should be run. In this way, subjects are acquainted with both the voting procedures and the types of impairments.

Finally, a mock test should be run, where a number of representative sequences are shown. The sequences should be different from those used in the test and they should be played in series without any interruption.

When the mock test is finished, the experimenter should mainly check that, in the case, when the test sequence is equal to the reference sequence, the evaluations of the subjects are close to 100. If instead the subjects declare seeing some differences, the experimenter should repeat both the explanation and the mock test [9].

3.2 PSNR - peak signal-to-noise ratio

The Peak Signal-to-Noise Ratio is the standard objective image quality measure. It is a ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. Because many signals have a very wide dynamic range, PSNR is usually expressed in terms of the logarithmic decibel scale. PSNR is most easily defined via the Mean Squared Error (MSE) for two images in resolution MxN pixels. f and f' are frames from the video sequence, where f image is considered reference and f' impaired approximation of the other. m and n are the actual pixels in the reference and impaired frame [7]. Root Mean Square Error (RMSE) calculates the “difference” between two images and is easily applied for each frame of the video.

For an MxN image, f is the original image and f' the impaired, RMSE can be calculated as:

$$RMSE = \sqrt{\frac{1}{M \times N} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} [f'(m,n) - f(m,n)]^2}$$  

(1)

PSNR is yet another extension based on RMSE and can be calculated as follows:

$$PSNR = 20 \log_{10} \left( \frac{255}{RMSE} \right)$$  

(2)

Typical values for PSNR are between 20 dB and 40 dB. Figure 2 shows noisy images at different typical PSNR values from 0 dB to 40 dB.

![Figure 2. Picture at different PSNR values: 40 dB, 30 dB, 20 dB, 10 dB and 0 dB.](image2)

Slika 2. Slika pri različnih vrednostih PSNR: 40 dB, 30 dB, 20 dB, 10 dB in 0 dB.

3.3 A Structural SIMilarity index method (SSIM)

This method is the state-of-the-art method for objective quality assessment and is structurally

![Figure 3. Images with the same mean squared error (MSE=142). A: original image; B: mean shifted image; C: contrast stretched image; D: blurred image; E: Joint Photographic Experts Group (JPEG) compressed image [7].](image3)

Slika 3. Prikaz slike z enakim srednjim kvadratičnim pogreškom (MSE=142). A: originalna slika; B: slika s spremenjeno srednjo vrednostjo; C: slika s povečanim kontrastom; D: razmazana slika; E: JPEG komprimirana slika [7].
dependent [7], [8]. Figure 3 show images that have the same MSE but have a different visual quality. This affects the SSIM method, so that it gives a different quality index than the PSNR method, where the dB value is the same as seen in Figure 3. During its long evolution and development processes, the human visual system (HVS) has been extensively exposed to the natural visual environment, and a variety of evidence has shown that the HVS is highly adapted to extract useful information from natural scenes [7]. Natural images are highly structured and the signal samples exhibit strong dependencies amongst themselves. These dependencies carry important information about the structures of objects in the visual scene. An image-quality metric that ignores such dependencies may fail to provide effective predictions of the image quality. Depending on how structural information and structural distortion are defined, there may be different ways to develop image quality assessment algorithms. The SSIM index is a specific implementation from the perspective of image formation. Figure 4 presents the structural diagram of SSIM. A Simplified SSIM equation is presented in the next few steps. First, the luminance of each reference signal x and the impaired signal y is estimated as the mean intensity, \( \mu_x \) is calculated in the same way but just with y signal components:

\[
\mu_x = \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i
\]  

(3)

We use the standard deviation as an estimate of signal contrast. An unbiased estimate in the discrete form is given:

\[
\sigma_x = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu_x)^2}
\]  

(4)

Similarly, \( \sigma_y \) can be estimated as:

\[
\sigma_y = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu_y)^2}
\]  

(5)

We introduce small constants C1, C2 in both the denominator and numerator. Finally, we combine these equations. The result is an index of image similarity measures, which we collectively refer to as SSIM. Indices between signals, x and y:

\[
SSIM(x, y) = \frac{(2\mu_x \mu_y + C_1)(2\sigma_{xy} + C_2)}{\mu_x^2 + \mu_y^2 + C_1}(\sigma_x^2 + \sigma_y^2 + C_2)^{\gamma}
\]  

(6)

4 Experimental results

Several video sequences were used for our assessment that exposes particular assessment problems but due to the size of the experiment only two are included in this paper. However, an evaluation result of a new system frequently depends heavily on the scene of the video sequence content. In general, it is essential to include critical sequences, especially when interpreting results, which are impossible to extrapolate from non critical sequences. The sequences “Foreman” and sequence from the movie “Matrix Revolutions” were chosen for our assessment experiment. The sequence “Foreman” is in resolution 352x288, the number of frames 300, color space YUV 4:2:0, source uncompressed progressive. The sequence Matrix is from DVD “Matrix Revolutions” including frames 120252 to 120552, coded in Motion Pictures Expert Group 2 (MPEG-2) known as H.262 [3], aspect ratio 2:40:1, NTSC, number of frames 300 of half the NTSC resolution. This sequence Matrix is very difficult to code. The main reasons for the difficulty are frequently brightness changes, very quick motion and frequent changes of scene.

<table>
<thead>
<tr>
<th>Average bit rate (kbit/s)</th>
<th>512</th>
<th>1024</th>
<th>2048</th>
<th>4096</th>
</tr>
</thead>
<tbody>
<tr>
<td>Video sequence Foreman</td>
<td>Grade</td>
<td>Grade</td>
<td>Grade</td>
<td>Grade</td>
</tr>
<tr>
<td>H.261</td>
<td>2.13</td>
<td>3.18</td>
<td>4.14</td>
<td>4.47</td>
</tr>
<tr>
<td>H.263+</td>
<td>3.04</td>
<td>3.64</td>
<td>4.32</td>
<td>4.53</td>
</tr>
<tr>
<td>H.264</td>
<td>3.82</td>
<td>4.35</td>
<td>4.56</td>
<td>4.82</td>
</tr>
<tr>
<td>Video sequence Matrix</td>
<td>Grade</td>
<td>Grade</td>
<td>Grade</td>
<td>Grade</td>
</tr>
<tr>
<td>H.261</td>
<td>1.12</td>
<td>1.82</td>
<td>3.25</td>
<td>3.67</td>
</tr>
<tr>
<td>H.263+</td>
<td>1.47</td>
<td>2.26</td>
<td>4.01</td>
<td>4.28</td>
</tr>
<tr>
<td>H.264</td>
<td>2.49</td>
<td>3.24</td>
<td>4.29</td>
<td>4.45</td>
</tr>
</tbody>
</table>

Average subjective assessment grades of the sequences “Foreman” and “Matrix” at average bit rates from 512 kbit/s to 4096 kbit/s.

4.1 SDSCE – subjective assessment

Table 1 presents results of assessment experiments made with a SDSCE method of subjective assessment in grades from 1 to 5, where 5 is the best possible grade. In this assessment, 19 subjects were included, which assessed under optimum condition as recommended in [9]. With H.261 the sequences were encoded with no annoying impairment over an average bit rate of 2000 kbit/s. The assessment results show that H.261 and H.263+ failed to code the “Matrix” sequence well under 2000 kbit/s and the impairments became annoying or even unwatchable. Left part of Figure 7 shows graphs of the SDSCE grades versus the bit rate for all codecs and sequences “Foreman” and “Matrix”.

Figure 4. Structural diagram of the image SIMilarity Measurement system [7].

Slika 4. Strukturni diagram ocejevanja indeksa podobnosti strukture [7].

Table 1

<table>
<thead>
<tr>
<th>SUBJECTIVE ASSESSMENT OF VIDEO SEQUENCES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average bit rate (kbit/s)</td>
</tr>
<tr>
<td>Video sequence Foreman</td>
</tr>
<tr>
<td>H.261</td>
</tr>
<tr>
<td>H.263+</td>
</tr>
<tr>
<td>H.264</td>
</tr>
<tr>
<td>Video sequence Matrix</td>
</tr>
<tr>
<td>H.261</td>
</tr>
<tr>
<td>H.263+</td>
</tr>
<tr>
<td>H.264</td>
</tr>
</tbody>
</table>

Povprečne ocene subjektivnega ocejevanja sekvenc “Foreman” in “Matrix” pri bitnih hitrostih 512 kbit/s do 4096 kbit/s.
Comparison of coded video sequences using the latest objective and subjective assessment methods

4.2 PSNR and SSIM objective assessment

The middle part of Figure 7, shows graphs of PSNR values versus average bit rate. The right part of Figure 7, shows graphs of the SSIM index versus the bit rate. Captures for the used codecs are presented in order to visually observe the differences in quality between different coders in Figure 8. The captures of video sequence “Foreman” is the frame with the number 157 and from sequence “Matrix” the frame with the number 74.

Results of the objective measurement using the PSNR and SSIM methods show that, for the test sequence “Foreman” coded over approximately average bit rates of the 1024 kbit/s with, H261, and H263+, there is no perceptual gain in quality. For complex sequence “Matrix” coded over 2048kbit/s, there is no perceptual gain in quality. This values for average bit rates are very interesting, because over it the objective methods are more suitable for quality assessment as subjective methods because human visual system is unable to perceive very small degradations. This breaking point varies and depends of video resolution, used codec and the scenes in the video.

At high-bit rates, both objective methods gave similar results. At middle and low bit rates, the SSIM method is more in accordance with the SDSCE method than the PSNR method. Finally, we can see that codecs H.261, H.263, and H.264 give similar results.
and H.263+ are quite good for the video sequence “Foreman” but are far from capabilities of the new generation codec H.264. The very complex sequence “Matrix” shows that old codecs H.261, H.263+ were unable to efficiently encode it in the quality-bit rate sense, because of the intensive motions and high-speed changes in the scene. All sequences coded with H.264 had low loss in quality for the human visual system. Codec was used in the main profile with single pass encoding. The new H.264 codec has high encoding efficiency, but needs four times more processing time than H.263+.

5 Conclusion

Our experiments included the latest objective and subjective quality assessment of digital video sequences encoded by using the standard older H.261, H.263+ and newer H.264 codec. Experiments show clearly that when assessing a high quality, high definition and high frame rate video sequence, the human visual system fails to assess small quality degradation and for which reason assessment must be done using objective methods.

In our experiments, the PSNR method was used for being a well-known and widely used method. The newest SSIM method was also used for obtaining more correlated results with human visual assessment, because the structural information from the images is considered in evaluation.

The subjective SDSCE method is better for quality assessment at low bit rates, for H.264 under 2048 kbit/s. The SSIM method at these bit rates more agree with the subjective grades than the PSNR method.

The codec H.264 outperformed both the older standard codecs H.261 and H.263+. The codec encodes video almost at half of the size of that obtained by the H.263+ codec and consumes four times more processing time. The new generation codec H.264 has therefore a great potential for future use, from new mobile phones to HDTV video and even digital cinema. However, the capabilities of this codec are at this stage not yet fully exploited because of its low processing power currently available.

6 References


Marko Hebar received his B. Sc. in Electrical Engineering from the University of Maribor in 2003. He is currently a Ph. D. student. His research focus is on image, video coding and remote sensing. He is a student member of IEEE.

Peter Planinšič received his B. S., M. S. and Ph. D. degrees from the University of Maribor, Slovenia, in 1979, 1991, and 2000, respectively. From 1981 to 1986, he was a Staff Scientist at Elektrokovina Maribor. Since 1986, he has been working at the University of Maribor, as a Docent for Electrical and Computer Engineering. His research interests include digital signal and image processing, image compression and communication, fuzzy-neural networks applied to the image quality and remote control. He is an author and co-author of many scientific papers and research project reports and is a member of SPIE and IEEE.