Quality assessment of D10 and DV25 video codecs for broadcasting purposes

Ljubomir Jovanov, Ewout Vansteenkiste,
Wilfried Philips
Image Processing and Interpretation Group (IPI-IBBT)
Ghent University
Sint-Pietersnieuwstraat 41, 9000 Ghent
ervsteen@telin.ugent.be

Wim Ermens, Tom Beckers
VRT-medialab
Gaston Crommenlaan 10 (bus 101)
9050 Ledeberg
wim.ermens@vrt.be

Abstract

In this study we investigate the image quality of two video codec standards (D10 and DV25) for broadcasting purposes. Different video sequences from the Flemish public broadcasting institute (VRT) archive were produced in Betacam format and subsequently compressed following the two codecs. Then, two instrumental measures being the Peak Signal to Noise Ratio (PSNR) and the Structural SImilarity Measure (SSIM) were used to assess the compressed video quality quantitatively. A psychovisual experiment based on the Multidimensional Scaling Framework (MDS) was set-up to assess the video quality qualitatively. Both approaches were compared and indicated unanonymously the preference for the D10 codec over the DV25 codec.

1 Introduction

We are entering the era of digital television. With HD-ready and (full)-HD standards finding their ways to the consumer market national broadcasting companies are forced to both innovate and adapt. Adapting prospectively is done easily by switching to HD-resolution acquisition, production and broadcast equipment. Retrospectively however, broadcasting companies possess huge archives of film and tape material that also have to be converted and preferably compressed (for conservation).
This study is part of the DIVA project (DIgital Vrt Archive) launched by the Flemish public broadcasting institute (VRT) and aims at defining the best video codec for the broadcasting of converted digital Betacam movie material. The Betacam film material was converted at VRT internally. Subsequently, conforming to the project protocol, two well-known codecs were selected to compress the converted material for broadcasting, the SMPTE D10 and DV25 standards for television. We will not elaborate on the details of these codecs in this paper but would like to refer to the following references for technical specificities [4, 3].
Usually, image/video quality is assessed quantitatively by what are called instrumental quality measures such as PSNR or (video) SSIM. However, it has been shown that those measures do not always correspond to human visual quality perception [2]. Thus, in this study we tried to correlate our quantitative findings to a psychovisual experiment we set up. We will show that in the case of digitized Betacam material, both the quantitative and qualitative approach favor the D10 codec.
The paper is organized as follows: We start with explaining how our test material was generated. Subsequently, our quantitative and qualitative approaches to image quality are explained. We will then present our results and end with a discussion of those and some conclusions.

2 Experimental Data

As mentioned in the introduction, our goal was to assess the quality of encoded Betacam material. All image data resided from the VRT archive. Betacam is a tape based digital video format, produced by Sony, and is commonly accepted as a high qualtity format.
The actual file conversion, from the Sony tape to the digital archive, is done through an Avid Media Composer at the VRT-site. The tape material is passed on to the Clipster machine as an uncompressed video stream over an SDI connection and saved there. Subsequently, the Avid Media Composer generates the two compressed files. Thus, in summary we have:

- Digibeta = material copied to digital Betacam: our "source" material
- D10 = source material converted to D10, 50Mbit/s compression, 4:2:2 color resolution

Figure 1. still from the non-compressed "Haspengouw" sequence

Figure 2. still from the D10 compressed "Haspengouw" sequence

- DV25 = source material converted to DV25, 25Mbit/s compression, 4:2:0 color resolution.

Finally, the compressed sources are again passed on to a Clipster machine over SDI to assure consistency in file formatting.A sample of the Digibeta source material, D10 and DV25 image formats (for one still image) can be seen in Fig. 1, 2, 3. As all formats are broadcast quality video production formats, it is clear already that the 3 formats resemble each other a lot.

3 Quantitative Analysis

Two well-known full-reference instrumental similarity measures are used in this study, the PSNR and SSIM [5]. The PSNR is the most straightforward pixel comparison measure, the SSIM is designed to be more true to the human visual system.
Suppose I is a reference (greyscale) image and I' an either noisy, filtered or compressed version, then the PSNR is

Figure 3. still from the DV25 compressed "Haspengouw" sequence

computed as

$$PSNR(I,I') = 20\log\frac{255}{\sqrt{\sum_x\sum_y(f_{I'}(x,y)-f_I(x,y))^2}} \quad (1)$$

where $f_I(x,y)$ corresponds to the grey value of pixel in I at position (x,y). PSNR is expressed in dB and, as can easily be seen from the formula, the higher the PSNR the higher the pixel-to-pixel correspondence between two images.

The SSIM separates the task of similarity measurement between two (greyscale) images I, I' into three comparisons: luminance, contrast and structure combined as

$$SSIM(I,I') = [\mathbf{l}(I,I')]^\alpha.[\mathbf{c}(I,I')]^\beta.[\mathbf{s}(I,I')]^\gamma \quad (2)$$

The luminance component $\mathbf{l}(I,I')$ is defined as

$$\mathbf{l}(I,I') = \frac{2\mu_x\mu_y + C_1}{\mu_x^2+\mu_y^2+C_1}, \quad (3)$$

where $\mu_{x,y}$ denote the mean intensity over row and column pixel values and C_1 denotes a constant to avoid instability when $\mu_x^2+\mu_y^2$ approaches zero. Usually $C_1 = K_1L$ with L the dynamic range of the image and $K_1 < 1$.
The contrast component $\mathbf{c}(I,I')$ is defined as

$$\mathbf{c}(I,I') = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2+\sigma_y^2+C_2}, \quad (4)$$

where $\sigma_{x,y}$ denote the standard deviation of row and column pixel values and C_2 denotes a constast to avoid instability when $\sigma_x^2+\sigma_y^2$ approaches zero. Usually $C_2 = K_2L$ with L the dynamic range of the image and again $K_2 < 1$.
Finally, the structure component $\mathbf{s}(I,I')$ is defined as

$$\mathbf{s}(I,I') = \frac{\sigma_{xy} + C_3}{\sigma_x^2\sigma_y^2 + C_3} \quad (5)$$

Figure 4. still from the“Jaar 1963” sequence

Figure 5. still from the“Boeketje” sequence

Figure 6. still from the “Haspengouw” sequence

with σ_{xy} the row and column correlation cofficient

$$\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y). \qquad (6)$$

C_3 is again a stabilization constant, chosen as $C_3 = C_2/2$. The $\alpha, \beta, \gamma > 0$ parameters to adjust the relative importance of each component our all set to 1. K_1 and K_2 were set to 0.01. Note that the SSIM ranges between 0 and 1 where total image resemblance results in an SSIM of 1.

4 Qualitative Analysis

In essence, in a psycho-visual experiment we try to quantify subjective human sensations to specific stimuli, as opposed to quantitative instrumental measures aimed at defining all of those in one objective number. Usually, the following terminology is applied: the people involved in an experiment are called the subjects or observers. The sensations or quality criteria they judge and score are called the attributes. The objects they actually score the sensations on are called the stimuli, i.e. the codecs in our case. The way these stimuli are scored and processed is called the methodology.

4.1 Stimuli

Three digitized Betacam sequences of 30 seconds each are used in this experiment: “Jaar 1963”, “Boeketje” and “Haspengouw”, see Fig. 4,5,6. The scene content selection was done at the VRT and based on a range of different detail information they wanted to be included in the test (e.g. fast movement, slow movement, spatial details vs. more homogeneous backgrounds, fast scene changes/cuts vs. gradual fading out).

4.2 Subjects

Thirteen viewers took part in the experiment. All subjects were experienced viewers either working at the VRT site or Ghent University image processing department. All images were shown on a professional Panasonic display in split screen mode, calibrated conforming to ITU-regulations [1]. The viewing distance was fixed at about 3 times the screen height.

4.3 Methodology

A double-stimulus experiment, where two sequences are shown in split-screen mode simultaneously, was performed. In the experiment the viewers were asked to score the similarity of the video sequences, on a discrete scale of ”0” to ”5”, as well as the preference of one sequence over the other in overall quality, on a scale of ”-2” to ”+2” (”-2” meaning a preference for the leftmost sequence, ”+2” meaning a preference for the rightmost image). All possible combinations for the same scene were presented, with each sequence once on the left-hand side on screen and once on the right-hand side, resulting in 18 pairs to score. All viewers had to go through a training phase where the range of possible image distortions and degradations were shown in order to tune their scaling. The entire experiment, training included, took about 30 minutes per subject on average.

Once the scores were gathered, a multidimensional scaling (MDS) framework was applied to process the data. We

will only explain the general idea about this framework in this paper, for more detailed information we refer to [2]. In essence, what happens is that from the similarity data gathered, through an iterative maximum likelihood (ML) procedure, the input samples $i = 1, 2, 3$ (source material, D10 and DV25) are represented by points $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ in a (multidimensional) perceptual space and arranged in such a way that the distance between the points in that space corresponds linearly to the perceived image similarity. Subsequently, the preference data is treated in a similar way, resulting in an attribute axis in the perceptual space in such a way that the orthogonal projection of the points onto the attribute axis results in the ordering of the stimuli conforming to the attribute. In our experiment, we are determining and projecting onto an overall preference in quality-axis.

5 Results

In Fig. 7, 8, 9 the frame by frame PSNR values are computed over the luminance channel for the 3 sequences under investigation. The Fig. 10, 11, 12 represent the SSIM values plotted frame by frame and calculated on the luminance channel as denoted in Section 3. Note also that since the SSIM is usually applied locally it is calculated in moving windows of 8by8-pixel blocks and the average per frame SSIM is plotted.

In Fig. 13 the psychovisual MDS space is plotted as obtained from the psycho-visual experiment explained in Section 4. On the X-axis the different file formats (codecs) are shown, on the Y-axis the projection on the perceived image quality axis is shown.

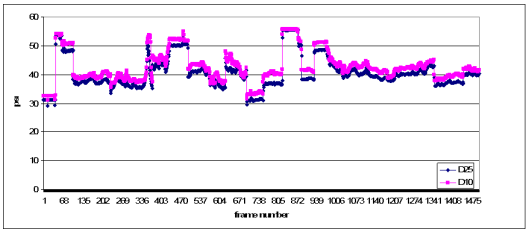

Figure 7. PSNR (dB) for the "Jaar 1963" sequence

6 Discussion

As can be seen from Fig. 7, 8, 9 as well as from Fig. 10, 11, 12 the quantititive analysis clearly indicates that D10 is always preferred over DV25. The difference expressed in the PSNR values, where on average we obtain a significant difference of about 1dB for "Jaar 1963",

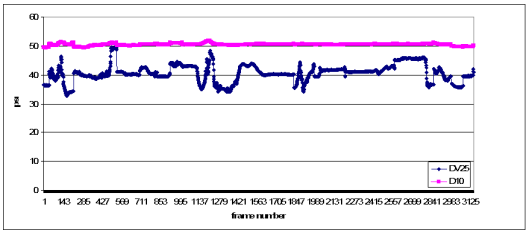

Figure 8. PSNR (dB) for the"Boeketje" sequence

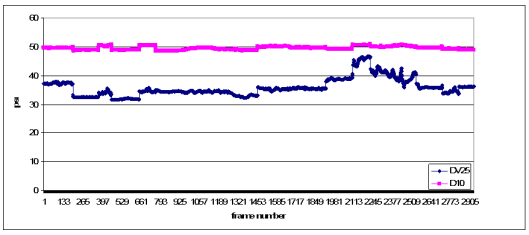

Figure 9. PSNR (dB) for the "Haspengouw" sequence

10dB for "Boeketje" and 13dB for 'Haspengouw", suggest that there would be a huge difference between the codecs. However, when visually inspecting the images this is not at all the case. This indicates that possibly a mere pixel to pixel comparison in the classical is not the way to go. Consequently, we believe the more subtle difference is represented more reliabley in the SSIM scores with an average difference of 0.004. Turning to the qualitative comparison, Fig. 13 clearly supports the quantitative results. Although the numbers on the Y-axis do not have any absolute meaning in terms of units of quality, we clearly perceive a consistent ranking in quality for the different file formats for all 3 scenes.

A statistical analysis of the qualitative data also resulted in the confidence intervals plotted as the error bars and corresponding to the 95% confidence intervals over all subjects

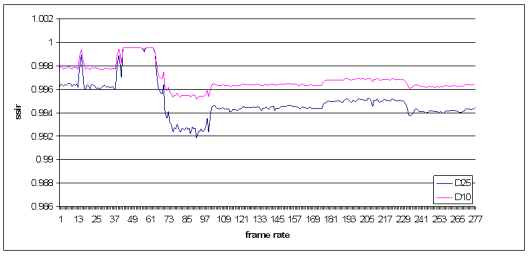

Figure 10. SSIM for the "Jaar 1963" sequence

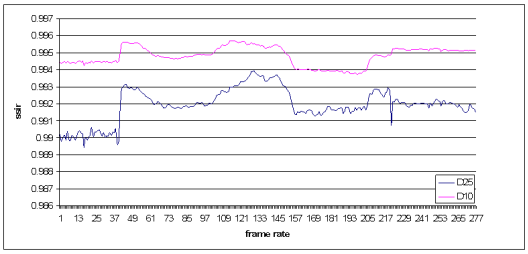

Figure 11. SSIM for the“Boeketje” sequence

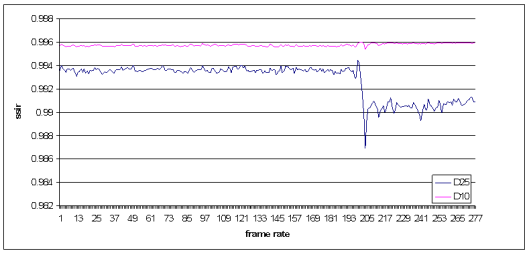

Figure 12. SSIM for the “Haspengouw” sequence

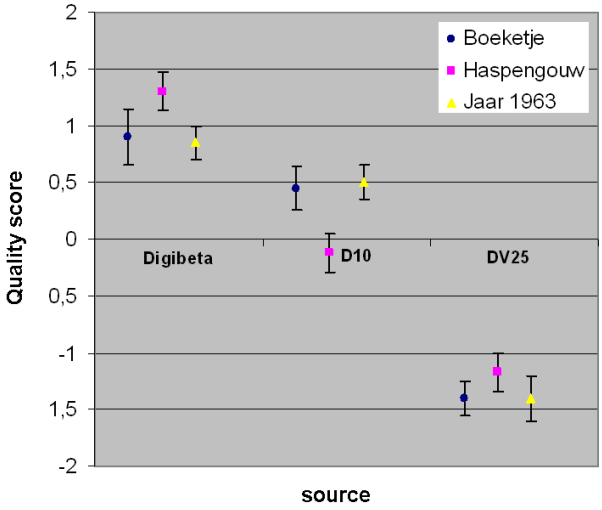

Figure 13. Psychovisual MDS space for the three sequences. On the X-axis the different codecs. On the Y-axis the perceived image quality. The error bars on the plot indicate the 95% confidence intervals over all subjects involved in the experiment.

participating in the experiments. In other words, the small error bars indicate the the position of the points in the perceptual space are highly reliable and that all viewers did agree.

7 Conclusion

In this study we investigated the image quality of the state of the art video codecs for their possible application in digitizing video archives. We were able to show both in an quantitative and qualitative way that the D10 codec outperforms the DV25 codec, yet also that we have to be careful with interpreting the results of standard quantitative comparisons.
Based on this research, the VRT has recently started the final conversion of the entire Betacam archive. Note that, although we only presented results here for Betacam data we are also performing similar experiments on 16mm and 35mm film material within the same DIVA project.

References

[1] ITU. Recommendation bt.500-11, methodology for the subjective assessment of the quality of television pictures. Technical report, ITU, Geneva, 2002.

[2] J.-B. Martens. *Image Technology Design*, chapter Psychophysical measurement and modelling of image quality. Springer, 2003.

[3] S. of Motion Pictures and T. Engineers. 6.35-mm type d-7 compentent format - video compression at 25 mb/s and 50 mbs - 525/60 and 625/50. Technical report, SMPTWE, 2002.

[4] S. of Motion Pictures and T. Engineers. Type d-10 stream specifications - mpeg-2 4:2:2p@ml for 525/60 and 625/50. Technical report, SMPTWE, 2006.

[5] Z. Wang, A. Bovik, and E. Sheikh, H. abd Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE transactions on images processing*, 13(4):600–12, 2004.