# Analysis of Coding Tools and Improvement of Text Readability for Screen Content

Holger Meuel, Julia Schmidt, Marco Munderloh, Jörn Ostermann
Institut für Informationsverarbeitung, Gottfried Wilhelm Leibniz Universität Hannover, Hannover, Germany
Email: *{lastname}*@tnt.uni-hannover.de

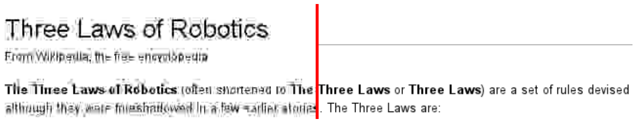
*Abstract*—**Current video coding standards perform well for video sequences captured by a real camera. The aperture of the camera's optical system smooths the content and attenuates higher frequencies. New application scenarios, enabled by the growing number of high bit rate internet gateways, however, make it necessary to take a closer look at the efficiency of such standards in handling artificial content. Remote desktop applications for example often include text parts. As a consequence, these content types contain sharp edges or high frequencies, which are considered less important in natural video and are therefore treated less carefully. The frequent result is an increased occurrence of artefacts or the loss of information that is actually important to the user. This paper gives an analysis of such artificially created video sequences, evaluates the performance of current coding tools for this type of content and proposes a simple, yet effective way to maintain readability of text within video material using only well considered encoder control and without the need of large additional modules.**

## I. Introduction

Most of the prevailing video coding standards perform well for natural camera-captured sequences, since the aperture of the optical system smooths the content by attenuating higher frequencies. However, sharp edges, which are present in most text applications, produce artefacts which impair the user experience or distinctly increase the data rate of the coded video sequence. An example for these artefacts is displayed in Fig. 1. Other properties of artificial as opposed to natural content are the absence of noise in most cases as well as lens distortion, as there is no camera involved in the image capturing process that could add such effects. Video sequences containing the described characteristics are often created by capturing computer desktop screens. Therefore these kind of sequences are called screen content (SC) sequences.

Text is a main element in many SC applications, which is why the focus of this paper lies on the improvement of the coding of these regions. As stated in [1], low complexity tools should be studied to enable good performance for SC with negligible harm to other test sequences. Our approach selects regions containing text via a simple edge detection, creates a region of interest (ROI) map and evaluates it to code the text-heavy regions using a lower Quantisation Parameter (QP) compared to the rest of the frame. This enhances the quality of the text areas. The ROI coding control used here is based on the model introduced in [10].

In the following section an introduction into state of the art methods for screen content coding (SCC) and coding control is given. In Section II an analysis of SC properties and coding tools helps to understand the particulars at hand, in Section III our algorithm is explained, followed by experimental results in Section IV and conclusions in Section V.



Three Laws of Robotics

From Wikipedia, the free encyclopedia

**The Three Laws of Robotics** (often shortened to **The Three Laws** or **Three Laws**) are a set of rules devised although they were foreshadowed in a few earlier stories. The Three Laws are:

Figure 1: Transform based coding of text content. The part left of the red bar is lossy, the part right of it lossless coded.

### A. State of the art

For text and computer generated graphics content, the signal in the spatial domain is sparser than in the transform domain. Residual Scalar Quantization (RSQ) is an approach to utilise this attribute [8] by not transforming the directional prediction residual in the frequency domain but instead quantising it in the spatial domain. Another way of exploiting the SC characteristics, in this case the reduced colour spectrum, is clustering to find the base colours of a block, which are stored in a table. Every sample of a block is quantised to its nearest base colour. The indices of the quantised values are saved in an index map [4][8]. To further enhance these methods, motion compensation is added in [7] and a Lempel-Ziv-Markov algorithm is applied to exploit repetitive patterns in text areas. In [9] the image blocks are classified into categories dependant on their type of content and are treated accordingly. Since in all of those publications additional modes are added for special image contents like text or graphics, a lot of supplementary encoder complexity is introduced. The most important difference is that the structure of the resulting bit stream is incompatible to AVC and thus needs special decoders while the proposed approach is fully compatible with the AVC standard. However, both [4] and [9] applied their approach to still images. Virtual Network Computing (VNC), a cross-platform implementation of the Remote Framebuffer Protocol (RFB), contains a number of different encoding algorithms for SC [12] between which the user can choose.

The focus of this work especially lies in the improvement of text passages in SC applications by not adding too much complexity but instead using the existing architecture wisely. A simple solution for treating text would be to detect it and handle it differently compared to the rest of the image. These parts of the image are declared ROI, as in [5]. Such regions are coded using finer quantisation steps, which leads to less distortions and should therefore be better to read. As the rest of the image is not as important in applications that are envisioned for SCC the remaining macroblocks can be quantised coarser. With this redistribution of the available total bit rate we can maintain an overall bit rate similar to the one without text-detection as preprocessing in typical scenarios. However, if most of the macroblocks of the image are classified as text, the bit rate is elevated to the level of a lower QP (higher quality) encoding. In [3], a similar approach is applied with skin colour as a ROI. If enough bandwidth is available, there is no problem in coding screen content. Thus, we will focus on small bandwidth applications which are critical in terms of coding quality, especially in text blocks which have to be as readable as possible.

## II. Analysis

### *A. Analysis of screen content properties*

According to the JCT-VC standardisation activities [2] there are several typical scenarios for SCC. We categorise these application scenarios as follows:

1) Office applications (e.g. text/spreadsheet processing)
2) Text insertions into natural video (news tickers etc.)
3) Video Streaming for streaming services or online gaming
4) Hybrid video sequences containing an arbitrary mixture

Our approach mainly addresses the first two of the application scenarios identified as the remaining two contain only little text and therefore do not have a lot of potential for improvement in this area.

Typical office work scenarios match the motion model of translational motion perfectly, e.g. by moving application windows around the desktop or by scrolling through documents. We, however, have to deal with sharp edges introduced by letters and symbols which are designed to have high contrasts against the background. In natural image video coding it is assumed that higher frequencies are not that important to the human perception and therefore can be quantised strongly or even be neglected. The opposite is valid for screen content, assuming that an observer is more interested in text symbols than in an accurate reconstruction of the background.

There are many methods to distinguish between foreground (text) and background (colour gradients, images) that are well known to the Computer Vision and Image Processing Society. However, most of them are not acceptable for video encoding purposes for performance reasons. A detection method has to be fast and efficient in terms of detection rate.

### *B. Analysis of coding tools for screen content in* Advanced Video Coding *(AVC)*

Our experiments showed that the difference between the amount of data needed for I slices compared to that of P and B slices is much higher for SC than it is for natural video. While factors of approximately 20–1000 for I/P slices and 50–2000 for I/B slices is common for the latter, we get quotients between 500–10000 for I/P and 500–100000 for I/B for SC sequences. To sum up, the smaller the movement inside a sequence in general, the higher the data rate difference.

We investigated the usefulness of coding tools in the state-of-the-art AVC hybrid video codec [11] for the coding of screen content. In the list below we provide an overview of coding tools and explain whether they are appropriate for SCC or not.

- *Distance of Reference Frames:* Since most SC sequences of desktop applications contain relatively slow movements, having a lot of reference slices spread over time as wide as possible is beneficial for high efficiency coding. Of course there are application scenarios like *Random Access* which deny the use of high key intervals.
- *Hierarchical B slices*: Contrary to linear prediction modes, a hierarchical prediction structure is applied wherein B slices can be referenced by other B slices. For slowly changing content there is very little difference between frames. Adding a hierarchical B slice as a reference does not provide additional information but counts as one reference frame which lowers the maximal temporal prediction distance.
- *Number of B slices:* For low bit rates and slow changing content it is advisable to disable B slices completely. The data rate to encode the residuals of the P slices increases due to the larger temporal prediction distance. For small motion the increase is larger than the gain given by the introduction of B slices.
- *Adaptive Quantisation Parameters (QPs):* Our experiments showed that fixed QPs give the best results for sequences containing little changes. Signalling QP changes always introduces extra transmission costs which have to be compensated by a noticeably better visual result. This is hard to accomplish in low bit rate conditions.
- *Resolution of Motion Vectors (MV):* According to our experiments, a higher resolution of the MV increases the coding efficiency the same way that it does for camera captured natural video sequences.
- *Spatial and Temporal Direct Mode:* Short tests showed that the use of spatial and temporal *Direct Mode* stays the same in SC and camera captured sequences. Around 95–98 % of the Direct Mode coded blocks are better coded spatially, the remaining 2–5 % gain from selecting temporal Direct Mode.

## III. Methodology

A detection method for separating text and background has to be very efficient in terms of complexity and runtime, while the detection rate has to be high as well. In general the contrast between text and background is high. Our experiments showed that a *Canny Edge Detector* is fast and reliable for the detection of text areas. The text detection was performed as a preprocessing step whose result was used to determine which QP should be elevated and which should be decreased on a macroblock level in relation to a previously fixed basis QP. Finally we compared the text readability in subjective viewings. To assure a fair comparison between the reference implementation and our approach, all adaptive coding mechanisms were disabled. As reference we used the x264 software [13] with the same settings applied to our own modified

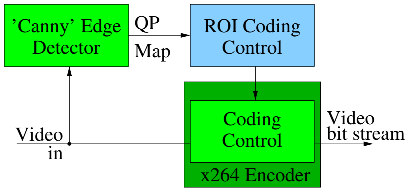


Figure 2: Block diagram of the coding concept

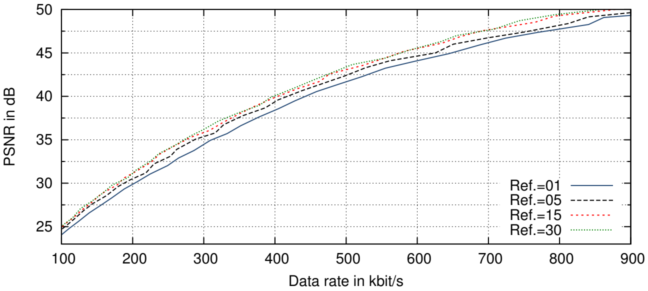


Figure 3: The RD diagram shows the coding performance for different numbers of reference slices for the sequence *Screen Capture Slide Editing*. Coding performance increases for up to 15 reference slices.

version of x264 which was extended for the possibility of controlling the QP adaptively (Fig. 2) for each macroblock (16 × 16) in the range of ±3.

### A. Evaluation Method

In general, common PSNR is not a perfect but an acceptable measure for automatic evaluation of image quality [6]. For natural video sequences typical PSNR values are in the range of 30 dB, which is bad quality with lots of image distortions, up to 45 dB, which indicates a near-to-perfect image reconstruction. For screen content, PSNR often leads to misunderstandable results as demonstrated in example frames from the JCT-VC test sequence *Slide Show*. They contain only a black screen with an arrow or a loading symbol. Since nearly the entire image is black, the subjective quality is good, but the resulting PSNR value of outstanding 62 dB for QP 37 does not mirror the real quality of details. Thus, PSNR measurement can be taken as an indicator of the quality but should not be overvalued, especially for SC and the analysis of text characters.

## IV. Experimental Validation

In this section research results for coding tools are evaluated on a PSNR basis, since only a direct comparison is of interest, despite this measure's shortcomings. The optimal number of reference slices for the sequence *Screen Capture Slide Editing* is shown in Fig. 3. Up to 15 reference slices increase the coding performance, a higher number of reference slices raises complexity without appreciable additional coding gain.

Of course different sequences have a different number of content changes within their progress but in general the less variation is within a sequence, the better disabling hierarchical B slices works. Fig. 4 shows the Rate Distortion (RD) curve for the two SC sequences *Slide Show* and *Screen Capture Slide Editing* using slowly changing content as an example. For very slow changes as in the latter (red and green curve) more coding gain can be achieved by disabling hierarchical B slices (≙ B slices=0) than when utilising the sequence *Slide Show* in which fast slide changes decrease the gain.

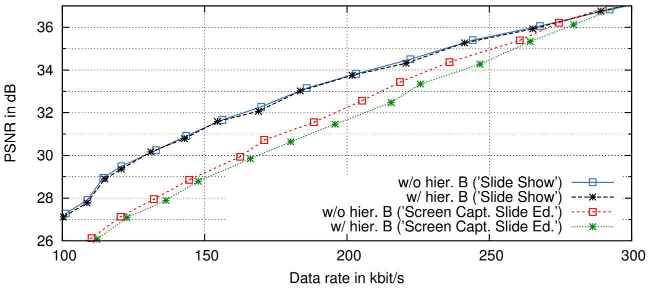


Figure 4: The RD diagram shows coding performance with and without hierarchical B slices for the sequences *Slide Show* and *Screen Capture Slide Editing*. For slowly changing content, disabling hierarchical B slices can improve coding efficiency at low bit rates.

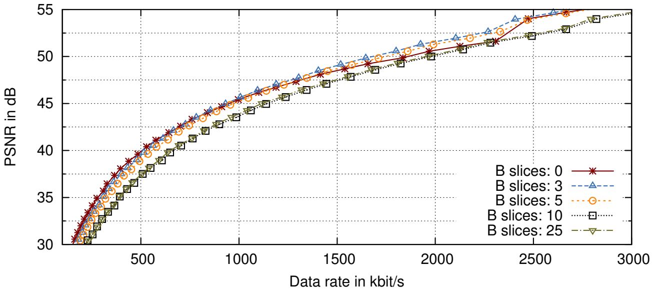


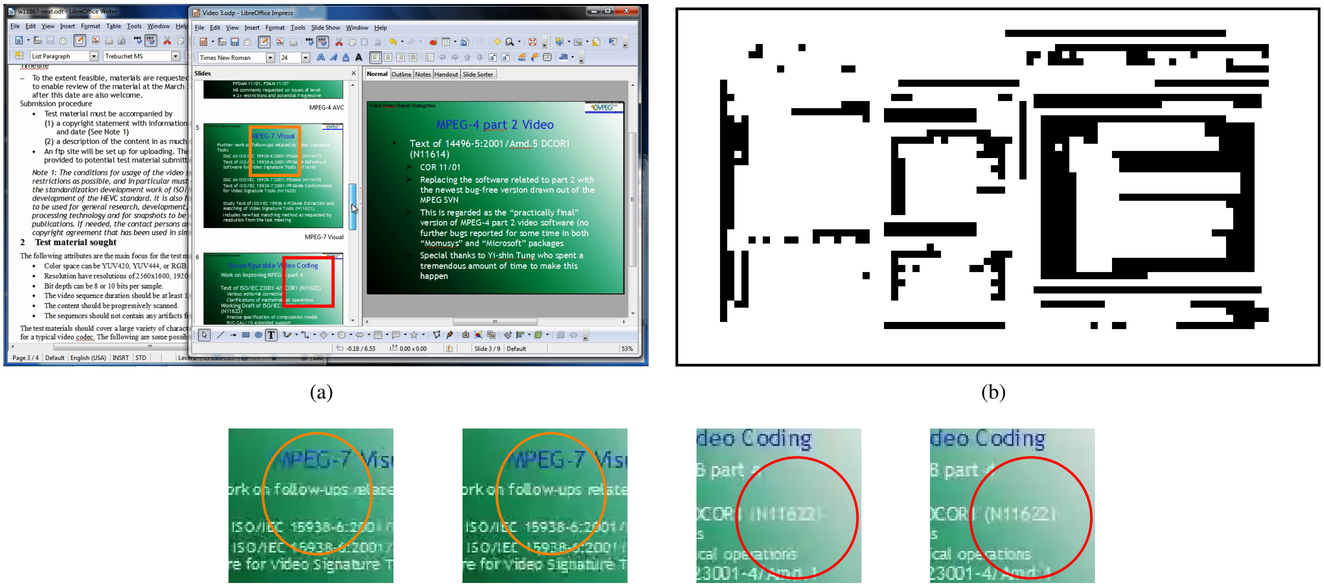
Figure 5: The RD diagram shows the coding performance for different numbers of B slices: For very low bit rates disabling B slices performs best, for most application scenarios 3 B slices are a good choice.

In our simulations for the optimal number of B slices we disabled adaptive B slice decision. At low bit rates below 1000 kbit/s, disabling B slices proved better for the coding efficiency of slowly changing screen content. For higher bit rates, it is advisable to set the number of B slices to around 3 (Fig. 5). While increasing this number to 5 can be beneficial for slowly changing sequences, 3 is a good number of B slices for the majority of sequences, and only performs slightly worse than 5 in the aforementioned situation. The optimal number of B slices is similar to the one in camera captured content.

We evaluated the dependency of the MV accuracy for the RD performance. While full pel MV accuracy is significantly worse (up to 1 dB in RD diagram) than any finer resolution, it is notable that half pel resolution (for macroblocks) has nearly no benefit when compared to quarter pel resolution of the MVs. This effect is mainly caused by the strong difference between I and P/B slices, since the predominant part of the overall bit rate is consumed by the I slice.

### A. Adaptive QP map encoding

As the PSNR indeed indicates if a test sequence is *better* or *worse* in terms of global image quality but cannot reflect the readability, which we considered as most important, we present our results from adaptive QP map coding in a subjective manner. Fig. 6 shows frame 44 from the JCT-VC test sequence *Screen Capture Slide Editing* [2]. This particular sequence has been chosen for demonstration because it contains lots of text and is hard to encode with general coding tools as our comparison shows. Of course the adaptive QP map also works


(a) (b)

(c) (d) (e) (f)

Figure 6: Frame 44 of the test sequence *Screen Capture Slide Editing* (1280 × 720). 6(a) Original Frame, 6(b) QP map used for coding (white=finer quantisation), 6(c) & 6(e) Frame after encoding with original *x264*, 6(d) & 6(f) Frame after coding/decoding using the QP map

for other sequences. With our edge based approach we only detect whether text is contained in a macroblock or not. If any text is detected in a macroblock, this block will be quantised finer (lower QP). Because of the increase in quality of single macroblocks, the overall bit rate might slightly increase. When compared to encoding with a lower global QP more than 5 % in bit rate can be saved in common scenarios while preserving the readability of text. A worst case scenario would be alternating macroblocks with and without text because of the additional signalling bits needed. This case, however, is very unusual in common application scenarios.

## V. Conclusion

In this paper we first analysed the general properties of screen content. In contrast to existing approaches which add a lot of computational load, we focused on using the architecture at hand wisely. We identified different application scenarios and analysed existing coding tools contained in AVC for their appropriateness when encoding screen content.

SCC experiments showed that in terms of coding efficiency it is useful to spread reference frames over time. Furthermore for low bit rates and slow changing content it is advisable to disable B slices completely, for bit rates above 1000 kbit/s 3 B slices performed best. Disabling hierarchical B slices performed worse in most cases and thus is not advisable.

To improve the subjective image quality with focus on the readability of text in SC sequences we developed a system to detect and control the encoding process of video sequences. The detection of text is performed by a *Canny Edge Detector*, whose output is analysed on a block basis in terms of high frequencies and corner characteristics. The resulting map is fed into a modified AVC coder to explicitly lower the QP in text areas and therefore increase readability. For sequences with very slowly changing contents, fixed QPs performs best. Although the bit rate is slightly increased when encoding some macroblocks in better quality than the reference, subjective quality of text is improved by the external adaptive QP control as readability tests showed. This method can save 5 % bit rate when compared to encoding the entire frame in a quality that guarantees good perception of text. Peculiarities of the *Canny Edge Detector* lead to the detection of window edges on top of the text blocks. To increase detection accuracy an additional text detector could lead to better results.

## References

[1] O. C. Au, J. Xu, H. Yu, *BoG report on Screen Content Coding (SCC)*, 4th JCT-VC Meeting, Daegu, Korea, Jan. 2011, Doc. JCTVC-D458.

[2] O. C. Au, J. Xu, H. Yu, *BoG report on Screen Content Coding (SCC)*, 6th JCT-VC Meeting, Torino (I), Jul. 2011, Doc. JCTVC-F771.

[3] P. Carillo, A. Osamoto, W. Jian, *Skin-tone Macroblock detection for Video coding*, 2011 IEEE Int. Symp. on Broadband Multim. Systems and Broadcast, pp. 1–4, Jun. 2011

[4] W. Ding, Y. Lu, F. Wu, *Enable Efficient Compound Image Compression in H.264/AVC Intra Coding*, Proc. IEEE Int. Conf. on Img. Proc., pp. 337–340, Oct. 2007

[5] N. Doulamis, A. Doulamis, D. Kalogeras, S. Kollias, *Low bit-rate coding of image sequences using adaptive regions of interest*, IEEE Trans. on Circ. and Systems for Video Technology, 8(8):928–934, Dec. 1998.

[6] A. Eden, *Eine Methode zur Messung der Bildqualität komprimiertere Videosequenzcen*, Dissertation, TU Braunschweig, 2010

[7] C. Lan, X. Peng, J. Xu, F. Wu, *Intra and inter coding tools for screen contents*, 5th JCT-VC meeting, Geneva (CH), Mar. 2011, Doc. JCTVC-E145.

[8] C. Lan, F. Wu, G. Shi, *Compress Compound Images in H.264/MPEG-4 AVC by Exploiting Spatial Correlation*, IEEE Trans. on Image Proc., Vol. 19, Issue 4, pp. 946–957, Mar. 15, 2010

[9] T. Lin, P. Hao, *Compound Image Compression for Real-Time Computer Screen Image Transmission*, IEEE Trans. on Image Proc., Vol. 14, No. 8, pp. 993–1005, Aug. 2005

[10] H. Meuel, M. Munderloh, J. Ostermann, *Low Bit Rate ROI Based Video Coding for HDTV Aerial Surveillance Video Sequences*, IEEE Conf. on Comp. Vis. and Pattern Rec. Workshop (CVPRW), pp. 13–20, Aug. 2011

[11] Recommendation ITU-T H.264, ISO/IEC 14496-10 (MPEG-4 Part 10): Adv. Video Coding (AVC) - 3rd Ed., ISO/IEC and ITU-T, Geneva (CH), Jul. 2004

[12] T. Richardson, *The RFB Protocol*, RealVNC Ltd, v3.8, May 17, 2004

[13] http://www.videolan.org