2012 IEEE International Conference on Multimedia and Expo Workshops

Video Description Length Guided Constant Quality Video Coding with Bitrate Constraint

Lei Yang
Google Inc.
1600 Amphitheatre Pkwy
Mountain View, CA, US
yangalalei@google.com

Debargha Mukherjee
Google Inc.
1600 Amphitheatre Pkwy
Mountain View, CA, US
debargha@google.com

Dapeng Wu
Electrical and Computer Engineering
University of Florida
Gainesville, FL, US
wu@ece.ufl.edu

***Abstract*—In this paper, we propose a new video encoding strategy — Video description length guided Constant Quality video coding with Bitrate Constraint (V-CQBC), for large scale video transcoding systems of video charing websites with varying unknown video contents. It provides smooth quality and saves bitrate and computation for transcoding millions of videos in both real time and batch mode. The new encoding strategy is based on the average bitrate-quality regression model and adapt to the encoded videos. Furthermore, three types of video description length (VDL), describing the video overall, spatial and temporal content complexity, are proposed to guide video coding. Experimental results show that the proposed coding strategy with saved computation could achieve better or similar RD performance than other coding strategies.**

***Keywords*-rate control; constant rate factor; multi-pass encoding; video description length; large scale video transcoding;**

I. Introduction

Videos have become an important part of human life in the digital age. The soaring number of videos demands efficient video compression, which is standardized in H.264/MPEG-4 Part 10 [1], [2] and the emerging H.265/HEVC [3], [4], [5]. Also the video sharing websites, such as YouTube and Vimeo, require to encode videos with the least bitrate, the least distortion and the least computational complexity with certain constraint. When real-time transcoding long videos with varying scenes, videos are chunked into pieces, parallelly transcoded and then concatenated together. Thus, simple and adaptive encoding strategies for smooth video quality and meeting bitrate constraint are desired.

There are many encoding strategies working for video compression, such as one-pass and multi-pass average bitrate encoding (ABR), constant bitrate encoding (CBR), constant quantizer encoding (CQP) and constant rate factor encoding (CRF) [6], [7]. These encoding strategies generally have the following properties, and serve for single objective.

ABR encoding strategy aims to achieve a target file size with file size error within the range of $\pm 10\%$ to meet network bandwidth constraint, but the quality of encoded video fluctuates due to the varying video content. CBR encoding strategy is designed for real-time streaming with constant bitrate. It has the fastest encoding speed but the lowest RD performance among all encoding strategies. CQP encoding strategy maintains a constant quantizer and compresses every frame the same amount by using the same quantization parameter (QP). It causes temporal perceptual quality fluctuation of encoded videos, especially when it uses large quantizers on videos with intensive scene change. CRF encoding strategy aims to constant visual quality with a constant rate factor (crf) with better perceptual performance and possible better RD performance than ABR encoding. But the output file size is unpredictable due to the varying video content. Therefore, it is hard to choose appropriate crf values to meet certain bitrate constraint of network or storage system for an arbitrary video. Besides, these conventional encoding strategies have varying performance on different videos, spend excessive resource for simple videos, and insufficient resource for complex videos in a large video pool. Unfortunately, they waste bitrate on simple videos, and may introduce blocky or blurring artifacts into complex videos.

To address these problems, we propose a new coding strategy—Constant Quality video coding with Bitrate Constraint (CQBC) based on the proposed bitrate-quality regression model to meet bitrate constraint and the least quality fluctuation at the same time. As far as we know, this encoding strategy is firstly proposed in the video coding literature. We also propose a RDC optimization method by properly assigning computation to each encoding pass of CQBC, and save 1/5 computation compared to other encoding strategies with better or similar RD performance. We proposed Video Description Length (VDL) by using relative encoded bitrate to describe video content complexity. Guided by VDL, CQBC in average saves computation to 3/4 of that of compared methods, saves bitrate by 2% on the test video set, and around 20% in real senario.

The paper is organized as follows. Section II gives a system overview of the paper. Then we study the bitrate-quality model in Section III. Based on the model, we propose the new coding strategy—CQBC and its optimization in Section IV. In Section V, three types of VDL are defined, and VDL guided Constant Quality video coding strategy

978-0-7695-4729-9/12 $26.00 © 2012 IEEE
DOI 10.1109/ICMEW.2012.70

with Bitrate Constraint (V-CQBC) is proposed in Section VI. Experimental results is shown in Section VII. Finally, we conclude this paper in Section VIII.

II. System Overview

The system overview of our paper is shown in Fig. 1. First,

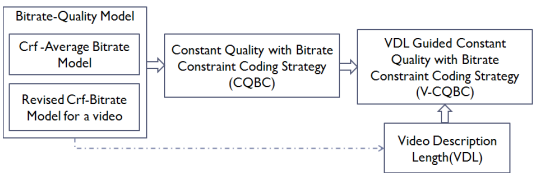


Figure 1. The system overview.

we study quality-bitrate model on a large multi-scene video corpus. Video quality is quantified by constant rate factor (crf) of x264 CRF encoding. By modeling crf-avgbitrate mapping, we can choose the appropriate crf which will generate bitrate close to the target bitrate in average. However, the videos have varying content. To alleviate the deviation of the actual bitrate of a specific video from the target bitrate, we propose a revised model to obtain a revised crf and encode that specific video with it to achieve the target bitrate with at most $\pm 10\%$ deviation.

Based on the bitrate-quality model, we propose the new coding strategy—CQBC. Its complexity could be reduced by the appropriate computation allocation among its multiple passes but still achieve similar or better RD performance.

Futhermore, we define three types of video description length (VDL) to describe the video overall, temporal and spatial content complexity. VDL could be obtained by a fast encoding algorithm, or from certain transcoding passes.

Accordingly, we use VDL to guide CQBC encoding, which is termed as V-CQBC. If the overall VDL of the current video is less than the average bitrate obtained from the model, then we can choose a relatively large crf value to encode the current video, which will shorten the encoding time as well as the number of iterations of CQBC algorithm, and vise versa. If the spatial VDL of the current video is larger than that of the reference, we can increase the complexity of encoding algorithm regarding spatial processing, and vise versa. Similarly, we tune the complexity of encoding algorithm regarding temporal processing according to the temporal VDL comparison.

III. Bitrate-Quality Model

A. Test Video Set

We build a large multi-scene video corpus based on the standard test videos [8], [9], [10], [11], with resolutions from QCIF to 1080P. Synthesized videos are generated by downsampling and randomly concatenating videos with at least four scenes to mimic real multi-scene videos. There are 400 test video sequences.

B. Crf-AvgBitrate Model

The average bitrate is a function of crf, spatial resolution, temporal resolution, when the encoding algorithm is fixed to be CRF with other coding parameters by default in x264. Due to the independence among these factors, the model by parameter separation is as following:

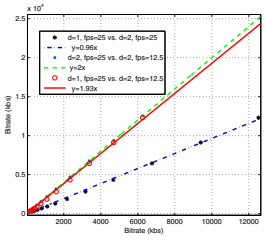
$$B = f(crf, M, T) = f_1(crf) \times f_2(M) \times f_3(T) \qquad (1)$$

where B is the average bitrate (kbps), M is the number of kilo pixels of Y component of a frame, T is the number of frames per second (fps).

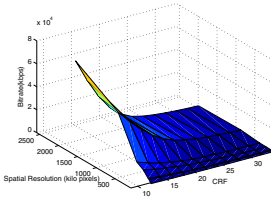
1) Crf-AvgBitrate Model of Temporal Resolution: The relationship between average bitrate and frame rate is modeled as a linear function as in (2), where parameter a includes the influence from spacial resolution and crf.

$$y = a \times T \qquad (2)$$

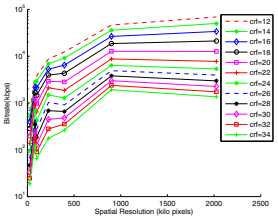
Since bitrate almost increases linearly with encoding frame rate (fps), i.e., $B_1/B_2 = fps_1/fps_2$ as shown in Fig. 2(a). The figure legend 'd' indicates downsampling rate, and 'fps' indicates encoding frame rate. For example, the points indicated by 'd=1, fps=25 vs. d=2, fps=12.5' have x coordinates denoting the average bitrate of videos downsampled by 2 and encoded by frame rate fps=12.5, and have y coordinates denoting the average bitrate of original videos encoded by frame rate fps=25. The points from right to left along each line in the figure are encoded with crf=12, 14, 16, $\cdots$, 34.



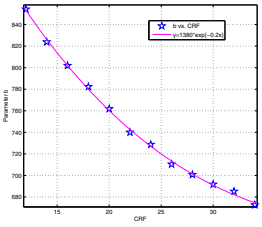
(a) Bitrate of videos with different temporal resolution



(b) Average bitrate with respect to crf and spatial resolution



(c) Average bitrate with respect to spatial resolution on all test videos



(d) Model parameter b in Eq. (3) as a function of crf.

Figure 2. Bitrate-Qaulity Modeling

2) Crf-AvgBitrate Model of Spatial Resolution: When the frame rate is fixed as 25 fps, the average bitrate with respect to crf and spatial resolution surface is shown in Fig. 2(b). From Fig. 2(b), we could see that bitrate is an approximate power function of the spatial resolution when fixing crf, and that bitrate is an approximate exponential function of crf when fixing the spatial resolution. For other frame rate, the bitrate is just a scaling of Fig. 2(b) along z axis by a factor of $fps/25$.

As shown in Fig. 2(c), the bitrate-spatial resolution polylines corresponding to different crfs are nearly parallel. The bitrate increasing rate is gradually decreasing along with the increase of the spatial resolution. Therefore, we propose to model the relationship between average bitrate and spatial resolution by power function

$$y = b \times M^c \tag{3}$$

where $0 < c < 1$ and is fitted to be 0.65, and b is a function of crf which is resolved when estimating the model between bitrate and crf when both temporal and spatial resolution are fixed.

3) Crf-AvgBitrate Model of Crf: Fixing spatial and temporal resolution, we use exponential function

$$y = m \times e^{n \times crf} \tag{4}$$

to model the relationship between average bitrate and crf, i.e. to model parameter b in Eq. (3) as a function of crf. The fitting curve is shown in Fig. 2(d), where parameter m is 1380, and n is -0.20. The fitting error is evaluated by SSE=540.3 and RMSE=7.351.

4) AvgBitrate as a Function of (T,M,crf): Based on the above modelling, the mapping between average bitrate B and (T, M, crf) could be evaluated by

$$\begin{aligned} B = f(T, M, crf) &= m \times e^{n \cdot crf} \times M^c \times \frac{T}{25} \\ &= 1380 \times e^{-0.2 crf} \times M^{0.65} \times \frac{T}{25} \end{aligned} \tag{5}$$

Accordingly, crf could be obtained from bitrate B.

$$\begin{aligned} crf &= f_1^{-1}\left(\frac{B}{f_2(M) \times f_3(T)}\right) \\ &= 5 \cdot \ln\left(\frac{55.2 \cdot M^{0.65} \cdot T}{B}\right) \end{aligned} \tag{6}$$

C. Revised Crf-Bitrate Model for A Video

For a specific video, the model is revised to be:

$$B = k \times f(crf, M, T) \tag{7}$$

where k is a revising factor determined from encoded videos.

IV. Constant Quality Encoding with Bitrate Constraint

A. New Coding Strategy

The algorithm 1 is a simple multi-pass encoding, similar to two-pass ABR encoding. The average number of encoding passes is 1.8.

Algorithm 1 Constant Quality Video Coding with Bitrate Constraint

/*Input: a video sequence, target bitrate B_t*/

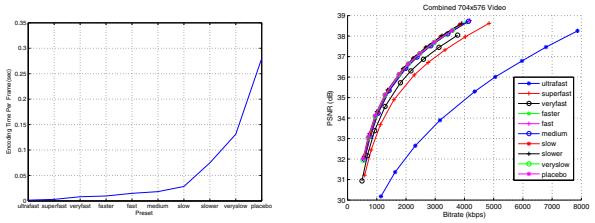
1: Find crf_t from the crf-avgbitrate model in Eq. (6) by substituting B with B_t;
2: Encode the video with crf_t, obtain the actual bitrate B_a;
3: Determine the revised model by (B_a, crf_t) pair;
4: Find crf_a from the revised model of Eq. (7) by substituting B with B_t;
5: Encode the video with crf_a, obtain the actual bitrate B'_a;
6: If B'_a does not fall in the range of $1 \pm 10\%$ of B_t, repeat from step 3 until convergence.

B. Algorithm Complexity Optimization

We evaluate the complexity of the coding strategy by encoding time per frame (sec). It is controlled by parameter 'preset' in x264, which takes ten values from 'ultrafast' to 'placebo', as shown in Fig. 3(a). For the proposed CQBC algorithm, the encoding time with 'preset=medium' is around 6 times of that with 'preset=superfast'. The RD performance increases along with the encoding algorithm complexity generally as shown in Fig. 3(b), and the RD performance is almost the same with 'fast' or even slower setting, except that RD performance with 'preset=ultrafast' deviates far from the average performance.

We set 'preset=superfaster' to the first pass and 'preset=faster' to the second pass of CQBC, the encoding time will be around 4/5 of that of one-pass ABR encoding. In this way, the encoding complexity of the proposed algorithm is lower than other encoding strategies, but still with higher or similar RD performance as shown in Fig. 6.

This RDC optimization method takes offline. By properly allocating computation to each encoding pass, multi-pass encoding could be RDC superior to one-pass encoding.



(a) Average encoding time per frame of CQBC when crf=26.

(b) RD performance of CQBC with respect to different presets

Figure 3. Performance of CQBC with each preset.

V. Video Description Length

The information about how many bitrates are needed to encode videos at certain quality reflects the video content

complexity. With this information, adaptive transcoding and RDC optimization is achievable.

Definition 1: The *Video Description Length* (VDL) is the bitrate needed to encode the video at certain quality.

We have *overall VDL* defined by absolute bitrate, and *temporal VDL* and *spatial VDL* defined by relative bitrate as following

Definition 2: The *overall VDL* is the actual bitrate of a video when it is encoded with 'crf=a, preset=superfast'.

Definition 3: The *temporal VDL* is the difference of the actual bitrate of a video when it is encoded with 'crf=a, preset=fast' and 'crf=a, preset=superfast'.

The difference of bitrate get rid of the spatial factor as much as possible with fixed crf.

Definition 4: The *spatial VDL* is the difference of the actual bitrate of a video when it is encoded with 'crf=a, preset=superfast' and 'crf=a+Δ, preset=superfast'.

The difference of bitrate get rid of the spatial factor as much as possible with fixed preset.

For video transcoding, VDL could guide us to choose the target bitrate, the target crf and encoding computation of transcoding to save bitrate and computation in terms of similar quality. It serves for transcoding video into multiple target formats, which include more than one hundred formats. We can compare the complexity of two videos with VDL, and determine the proper encoding parameters for the current video by referring to the existing reasonable encoding parameters of the reference video. A VDL reference table could be built when transcoding into one or two target formats, and then used to save bitrate and computation for transcoding into other target formats, and also in batch rerun transcoding.

VI. VDL Guided Constant Video Coding with Bitrate Constraint

We use VDL to guide CQBC encoding, which is termed as V-CQBC. With Algorithm 2, the average encoding time could be reduced to 3/4 of that of one-pass ABR encoding, and 2% of bitrate could be saved with video quality in terms of PSNR similar as before. Note that all the VDL information could be stored in a database as a basic information of videos, and reused repeatedly.

The average Algorithm 2's computation is saved to 3/4 of that of Algorithm 1. The bitrate is saved more than 2% on test videos. In real senario, the bitrate is saved around 20%.

VII. Experimental Results

A. Fitting Error Evaluation of Crf-AvgBitrate Model

The model in Eq. (5) is illustrated as a surface in Fig. 4.

Algorithm 2 VDL Guided Constant Quality Video Coding with Bitrate Constraint

/*Input: a video sequence, target bitrate B_t, VDL and encoding parameters of a standard video*/

1: Obtain the overall VDL, the temporal VDL and the spatial VDL of the input video;
2: If the overal VDL $< B_t$, set B_t = the overal VDL;
3: If the temporal VDL is less than the reference, reduce the temporal encoding algorithm complexity, and vise versa;
4: If the spatial VDL is less than the reference, reduce the spatial encoding algorithm complexity, and vise versa;
5: Call CQBC Algorithm 1.

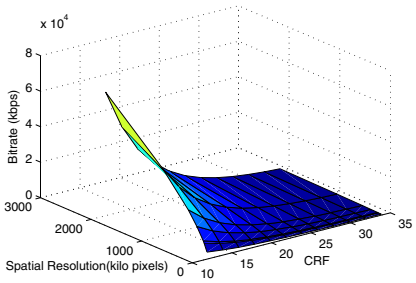


Figure 4. Fitting of mapping between bitrate and crf.

Table I
Relative fitting error on training and testing set.

Spatial Resolution	Training E_r	Testing E_r
176x144	0.43	0.33
352x288	0.39	0.45
352x240	0.41	0.37
640x360	0.22	0.25
704x576	0.17	0.16
1280x720	0.10	0.04
1920x1080	0.07	0.05

The relative fitting error is evaluated per spatial resolution by the equation below:

$$E_r(M) = \frac{\sum_{crf=12}^{34} \sum_{video_i \in \Omega_M} \frac{|B_i^a(crf,M) - B_i^e(crf,M)|}{B_i^a(crf,M)}}{|\Omega_M| \times 12} \quad (8)$$

M is the spatial resolution, Ω_M is the video set with spatial resolution M, $|\Omega_M|$ is the cardinality of Ω_M, E_r stands for the relative error, $B_i^a(crf, M)$ is the actual bitrate of the ith video with spatial resolution M encoded with crf, $B_i^e(M)$ stands for the bitrate of the ith video with spatial resolution M estimated from Eq. (5). The relative fitting error on training video set and testing video set are shown

in Table I. It shows that the relative fitting error is decreasing with spatial resolution increase, and that the relative fitting error on the testing videos is approximate to that on the training videos.

B. Evaluation of Revised Crf-Bitrate Model

For specific videos, the results are evaluated in the Table II. B_t is the target bitrate, B_a is the actual bitrate, which are in the unit of kbps, and k is the revising factor in Eq. (7).

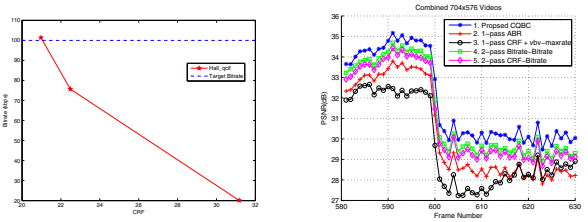
Table II
PERFORMANCE OF THE REVISED MODEL ON SPECIFIC VIDEOS

Videos	M	B_t	k	B_a
Mobile	176x144	100	0.50	91.53
Flower	352x288	300	0.75	293.95
Tennis	352x240	300	0.66	291.84
Parkrun	640x360	600	0.80	622.16
Harbour	704x576	1500	0.69	1457.02
Parkrun	1280x720	2500	1.05	2534.98
Pedestrian	1920x1080	3500	0.82	3313.23

From the Table II, we could see that if the coding performance on a specific video is far from the average coding performance on videos with the same spatial resolution, k will be away from 1, such as the first row in Table II. Otherwise, k will be close to 1 as the last two rows in Table II. The revised model in Eq. (7), promises the actual bitrate falls in the range of $(1 \pm 10\%)$ of the target bitrate.

C. Performance of CQBC

To encode a specific video towards the target bitrate, the number of encoding passes in Algorithm 1 is 1.8 in average in our experiments. A three-pass case is shown in Fig. 5(a) on video 'Hall_qcif'. The $(crf, bitrate)$ pairs are denoted by the points along the poly line from20kbps to 101.8kbps from right to left in Fig. 5(a). The crf values decrease in our algorithm to make the actual bitrate converge to the target bitrate 100 kbps.



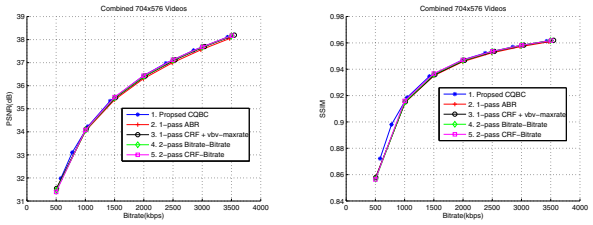
(a) Convergence of our coding algorithm 1 with multiple-pass case.

(b) PSNR fluctuation per frame.

Figure 5. CQBC encoding performace.

We compare PSNR performance of our encoding strategy with four encoding strategies, which all aim to achieve the target bitrate. They are

- 'proposed CQBC': proposed constant quality encoding with bitrate constraint;
- '1-pass ABR': one pass ABR encoding;
- '1-pass CRF + vbv-maxrate': one pass CRF encoding with a buffer size for bitrate constraint;
- '2-pass Bitrate-Bitrate': two pass ABR encoding;
- '2-pass CRF-Bitrate': two pass encoding with the first pass CRF encoding and the second pass ABR encoding.



(a) Bitrate(kbps) vs. PSNR(dB) of five coding strategies.

(b) Bitrate(kbps) vs. SSIM of five coding strategies.

Figure 6. PSNR and SSIM performance.

The Rate-Distortion performance of five coding strategies is shown in Fig. 6(a) and Fig. 6(b), in which distortion is evaluated by PSNR (dB) and SSIM respectively. The test video in these representive figures has 1200 frames including four scenes from sequences: city, crew, harbour and soccer, with spatial resolution 704x576. We can see that the 'proposed CQBC' encoding has the highest RD performance, and then '2-pass CRF-Bitrate' encoding, '2-pass Bitrate-Bitrate' encoding, '1-pass CRF + vbv-maxrate', and '1-pass ABR' encoding has the lowest RD performance. For the 704x576 video, the average PSNR gain of the 'proposed CQBC' relative to '1-pass ABR' is 0.15 dB and SSIM gain is 0.003 with the same bitrate. It holds similarly for other video resolution.

We also test five coding strategies all with the target bitrate 500kbps and other coding parameters by default. PSNR performance of each frame around scene change moment is shown in Fig. 5(b). The difference between the maximal PSNR and minimal PSNR of frames from 400 to 1200 of five coding strategies are 5.42 dB, 5.98 dB, 5.68 dB, 5.77 dB, 5.75 dB respectively. It indicates that the proposed CQBC encoding has the smallest PSNR change along the temporal direction of videos and the highest PSNR.

D. Evaluation of VDL

The content complexity order of single scene videos is shown in Table III in terms of overall VDL. The first video in each row is the most complex one, as we expected. The average overall VDL for each tested spatial resolution is: 123.3, 357.4, 570.5, 1587.1, 2820.8 and 4072.4 kbps respectively.

The temporal VDL comparison of videos with the single scene for each spatial resolution is shown in Table IV. The

average temporal VDL with respect to each tested spatial resolution is: 41.6, 85.2, 129.6, 149.9, 587.7 and 809.1 kbps respectively.

The spatial complexity evaluation of videos with single scene for each spatial resolution is shown in Table V. The average spatial VDL for each tested spatial resolution is: 30.3, 98.9, 167.4, 463.7, 1432.9 and 1058.2 kbps respectively.

Table III
THE OVERALL VDL COMPARISON.

Spatial Resolution	Overal Complexity Order
176x144	coastguard>mobile>container>suzie
352x288	flower>bus>tempete>foreman
352x240	garden>mobile>football>tennis
704x576	crew>harbour>soccer>city
1280x720	parkrun>stockholm>shields>mobcal
1920x1080	riverbed>tractor>pedestiran>station

Table IV
THE TEMPORAL VDL COMPARISON.

Spatial Resolution	Temporal Complexity Order
176x144	coastguard>container>suzie>mobile
352x288	flower>foreman>bus>tempete
352x240	tennis>mobile>football>garden
704x576	crew>soccer>harbour>city
1280x720	parkrun>mobcal>shields>stockholm
1920x1080	riverbed>pedestrian>tractor>station

Table V
THE SPATIAL VDL COMPARISON.

Spatial Resolution	Spatial Complexity Order
176x144	coastguard>mobile>container>suzie
352x288	flower>tempete>bus>foreman
352x240	mobile>garden>football>tennis
704x576	crew>harbour>soccer>city
1280x720	parkrun>stockholm>shields>mobcal
1920x1080	riverbed>tractor>pedestrian>station

VIII. CONCLUSION

In this paper, we investigated the bitrate-quality model on a large multi-scene video corpus, and proposed a new encoding strategy—constant quality video coding with bitrate constraint, which provides constant quality as well as satisfies bitrate constraint. Its computational complexity could be reduced by assigning small computations to each pass. Therefore, it had better rate-distortion-complexity (RDC) performance than other encoding strategies. We also proposed the overall video description length, temporal video description length and spatial video description length to describe video content complexity quickly, and used VDL to guide constant quality video coding with bitrate constraint. The algorithms saved computation and guaranteed the smoothest visual quality for parallelly video transcoding with video chunks as well as encoding whole videos with varying scenes.

The rate-distortion-complexity optimization of encoding strategies will be investigated in a quantified model further. The mapping between VDL and corresponding proper encoding parameters will be studied to assist VDL-guided video coding.

REFERENCES

[1] T. Wiegand, G. J. Sullivan, G. Bjntegaard, and A. Luthra, "Overview of the h.264/avc video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, 2003.

[2] I. Richardson, *H. 264 and MPEG-4 video compression: video coding for next-generation multimedia*. John Wiley & Sons Inc, 2003.

[3] R. Joshi, Y. Reznik, and M. Karczewicz, "Efficient large size transforms for high-performance video coding," in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, vol. 7798, 2010, p. 24.

[4] S. Vetrivel and K. Suba, "An overview of H. 26x series and its applications," *International Journal of Engineering Science and Technology*, vol. 2, pp. 4622–4631, 2010.

[5] D. Marpe, H. Schwarz *et al.*, "Video compression using nested quadtree structures, leaf merging, and improved techniques for motion representation and entropy coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, pp. 1676–1687, December 2010.

[6] L. Merritt and R. Vanam, "Improved rate control and motion estimation for h.264 encoder," in *ICIP (5)*, 2007, pp. 309–312.

[7] Z. Chen and K. N. Ngan, "Recent advances in rate control for video coding," *Image Commun.*, vol. 22, pp. 19–38, January 2007. [Online]. Available: http://portal.acm.org/citation.cfm?id=1224554.1224634

[8] "Qcif and cif sample videos," *http://trace.eas.asu.edu /yuv/.*

[9] "Hd sample videos," *ftp://ftp.ldv.e-technik.tu-muenchen.de/pub/.*

[10] "352x240 sample videos," *http://www.cipr.rpi.edu /resource/sequences/sif.html.*

[11] "704x576 sample videos," *ftp://ftp.tnt.uni-hannover.de /pub/svc/.*