Constant Frame Quality Control for H.264/AVC

Po-Chyi Su*, Ching-Yu Wu, Long-Wang Huang and Chia-Yang Chiou
Dept. of Computer Science and Information Engineering
National Central University, Jhongli, Taiwan
E-mail: *pochyisu@csie.ncu.edu.tw Tel/Fax: +886-3-4227151/+886-3-4222681

***Abstract*—A frame quality control mechanism for H.264/AVC is proposed in this research. The research objective is to ensure that a suitable Quantization Parameter (QP) can be assigned to each frame so that the target frame quality can be achieved. One application is consistently maintaining the frame quality during the encoding process to facilitate such applications of video archiving or surveillance. A single-parameter Distortion to Quantization (D-Q) model is derived by training a large number of frame blocks and this model parameter can be determined by the frame content. Given the target quality for a video frame, we can then select an appropriate QP according to the proposed D-Q model. The model refinement and QP adjustment of subsequent frames can be applied according to the coding results of previous data. Structural Similarity (SSIM) is chosen as the quality measurement to demonstrate the feasibility of the proposed framework.**

I. Introduction

H.264/AVC [1] is widely adopted in many applications requiring video compression due to its various advanced coding tools. Under the same quality constraint, the bit-rate saving of H.264/AVC is significant when compared with such predecessors as MPEG2 and MPEG4. It should be noted that the video frame quality is significantly affected by the Quantization Parameter (QP) assigned to each frame. Since the contents of video frames may be quite different, the video quality may fluctuate considerably and careless assignment of QP may result in serious distortion on certain frames. In such applications of video surveillance or archiving, we may require that the quality of each frame should be equally preserved well. In this research, we consider developing Distortion-Quantization (D-Q) model to help achieve the constant quality video coding. The measurement of quality has long been a research focus in video processing. Since the commonly used Peak Signal-to-Noise Ratio (PSNR) may not reflect the subjective quality well, many researchers [2], [3], [4], [5] tried to take human visual systems (HVS) into account to develop better evaluation methods. Structural SIMilarity index, SSIM [2], is a well-know one. SSIM of two contents x and y is defined as

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{({\mu_x}^2 + {\mu_y}^2 + C_1)({\sigma_x}^2 + {\sigma_y}^2 + C_2)}, \quad (1)$$

where μ_x (μ_y) is the local mean of x (y) and σ_x (σ_y) is the local standard deviation of x (y). σ_{xy} is the local correlation coefficient of x and y. C_1 and C_2 are small constants to avoid instability when the denominator is close to zero. Some researches showed that SSIM is more related to the human visual system and the coding algorithms explicitly using SSIM have also been proposed [6], [7]. The H.264/AVC reference code has included SSIM as one of its evaluation metrics.

Up to now, most of the existing work related to constant quality video coding adopted PSNR as the measurement. [8] is one of the early research to achieve this goal by encoding the video several times and employing Viterbi algorithm to pick a suitable QP for each frame. The encoding process is thus time-consuming and only acceptable in certain off-line applications. To attain more efficient quality control, D-Q and/or rate-distortion (R-D) models are proposed to facilitate the QP assignment. [9] determined the relationship between PSNR and QP to develop a Rate-Quantization (R-Q) model for effectively allocating bit budgets. [10] made use of Cauchy-density function to depict the distribution of AC coefficients of the Discrete Cosine Transform for developing an effective D-Q model. [11] used SATD (Sum of Absolute Hadamard Transform Differences) to determine the related parameters of a D-Q model, which can accurately predict the PSNR in intra coded frames. [12] assigned or adjusted the QP values according to the difference between the average PSNR of previously encoded frames and the target PSNR. If the difference is small, the QP of previous frame is used. [13] encoded the video twice and used the coded information of the first run as the reference to attain the constant quality coding.

Our ultimate goal is to propose a framework that can adopt more flexible quality measurements to develop a constant quality coding scheme. This paper is our first step for achieving a more general quality control scheme, and we choose to develop the D-Q model with the distortion measured by SSIM. Compared with the conventional quality metric PSNR, SSIM has a better representation of human's visual perception. Moreover, since the aforementioned existing research didn't provide reliable ways to estimate the model parameters, the commonly adopted approach is encoding the video several times, and then applying the linear/non-linear regression. The main contribution of this paper is that we can estimate the model parameter before the encoding process with some computationally efficient features. In other words, the proposed mechanism can help to achieve more efficient constant quality video coding. The rest of the paper is organized as follows. Sec. 2 will describe the model training of our proposed scheme and Sec. 3 presents the complete QP assignment procedure during the video coding. Sec. 4 demonstrates the experimental results, followed by the conclusion in Sec. 5.

II. D-Q Model

As we target at building a model that links the distortion and quantization/QP, the measurement of distortion has to be defined first. Since SSIM will be close to one if the contents to be compared are similar, we define the distortion D_{SSIM} as $1-SSIM$. It is observed that a power function can reasonably depict the relation between D_{SSIM} and the quantization in both the intra- and inter-coding. We employ the following function,

$$D_{SSIM} = \alpha \times QP^{\beta}, \tag{2}$$

with the two model parameters α and β. To verify the feasibility of this power function, we encode some test CIF videos, including Foreman, Coastguard, Container, Football, Mobile, Paris and Stefan, each with 100 frames. All the frames are intra coded with QP's ranging from 20 to 40 and the corresponding D_{SSIM} values are recorded. The curves from the collected data are matched with the above power function by the regression. We examine the accuracy of modeling by using the R^2 value, which is calculated by

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}, \tag{3}$$

where y_i is the actual value to be modeled, f_i is the corresponding resultant value calculated by the model, and $\bar{y}$ is the average value of all y_i. We can see that the R^2 values are all very close to one, which means that the proposed model can fit the actual D-Q curve accurately.

However, we list the parameters, α and β in Table I and we can find that the two values vary in each video. Existing work usually proposed to train some data in the same video or employed the data in the previously decoded frames for acquiring these parameters for the subsequent coding. The major disadvantage is that the quality fluctuation may be observed in the first few encoded frames without appropriate parameter setting or that multiple encoding processes are required. In addition, when the scene changes happen, the parameters have to be determined again or the performance will be affected seriously.

TABLE I
The Relationship between the distortion and QP.

Video	D_{SSIM} vs. QP	
	α	β
Foreman	2.3×10^{-6}	3.0
Coastguard	2.8×10^{-8}	4.4
Container	1.3×10^{-5}	2.6
Football	4.6×10^{-7}	3.5
Mobile	3.4×10^{-10}	5.4
Paris	2.9×10^{-8}	4.2
Stefan	1.7×10^{-10}	5.5

The objective of this research is to appropriately estimate these parameters by using a content-adaptive model. The first step is to collect more various data samples for further training. The frame will be divided into basic units. There are several

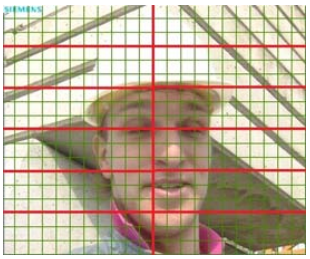


Fig. 1. Partition a frame into basic units.

choices for deciding the size of basic units, *e.g.*, the entire frame, a group of macroblocks (MB's) or a MB. Designing a frame model, *i.e.*, determining a QP value according to the feature representing the entire frame, sounds a reasonable and straightforward approach. This would be similar to what we have just shown that a feature representing the whole frame can help to determine α and β in Eq. (2) for the entire frame. However, we found that a slight model inaccuracy will result in a poor determination of QP. Using MB's directly for the model training should be more flexible. Nevertheless, according to our experience, when the unit size is too small, it will be difficult to determine a well-defined relationship between the content and the parameters. That is, the power function may not work well. An obvious example is that we may easily obtain a small block with uniform colors and encoding such blocks with different QP values may generate unexpected results. We may observe a larger distortion when a smaller QP is used. It should be noted that such blocks occupy a pretty large portion in common frames. In other words, there will be a large number of outliers in our training data. Training the model with so many "unusual" blocks will thus become challenging. Therefore, we choose to use a group of MB's as the basic unit. For a CIF video frame, we divide it into basic units as shown in Fig. 1. A unit contains 33 MB's so a frame consists of 12 basic units. By dividing the frame across the center as shown in Fig. 1, we can obtain blocks or basic units that contain more meaningful contents easily and the training process can be facilitated. That is, the number of outliers will be reduced dramatically. In addition, a larger number of units with more meaningful contents can certainly help the QP determination.

We first deal with the intra-coded frames. Since many frames in a video have similar content, we use still images for training, instead of video sequences. We use 200 images from Berkeley image database. Each image is scaled and cropped properly to the CIF frame size. These images are concatenated into a video, which is coded with various QP's. The quality distortion of each basic unit and the corresponding QP value are collected. It is observe that the relationship between the distortion and QP shown in Eq. (2) still holds. Nevertheless, we found that there exists a linear relationship between $ln(\alpha)$ and β as shown in Fig. 2. The R^2 values of the linear relationship are both as high as 0.99. The fact indicates that Eq. (2) can be reduced to only one variable. For I frames,

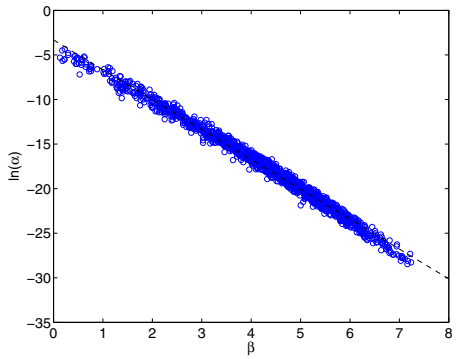


Fig. 2. The relation between $ln(\alpha)$ and β for using SSIM as the measurement in I frames.

the D-Q model can thus be expressed as

$$D^I_{SSIM} = e^{-3.35\beta - 3.32} \times QP^{\beta}. \quad (4)$$

In fact, according to our tests, the similar relation can be found in P frames and the data can be fitted well by

$$D^P_{SSIM} = e^{-3.48\beta - 2.55} \times QP^{\beta}. \quad (5)$$

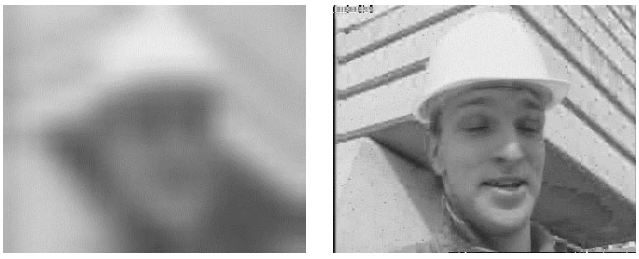
The R^2 values in P frames can also reach 0.99 in the SSIM distortion measurements.

The next step is to seek an efficient way to choose suitable β for a basic unit. It is worth noting that β is content-related. According to our observations, if the content will be affected by the lossy coding more easily, the value of β will be larger. On the other hand, for the unit with relatively more uniform content, β will be quite small. Therefore, we would like to predict the effects of compression on the content so that a reasonably good β can be selected. One way to achieve this is to encode the frame with different QP's to observe the curve but it may be computationally prohibited. In other words, this "pre-processing" has to be efficient to avoid considerably increasing the computational load of video coding. We adopt the follow strategy. The pre-processing or, in fact, a process of distortion, is applied on the input frame and then the selected quality measurement will be used to evaluate the degradation of these distorted versions. That is, we make use of these degradation measurements to help us select a suitable β.

Again, we collect training data for coding with different QP's to determine β and pre-process these training data to obtain the distortions. By examining β and the degradations, we would like to know whether such a solid relation exists. After various trials, the preprocessing we consider right now includes two parts: resizing and singular value decomposition (SVD). The resizing process quickly removes the high-frequency textures. We simply calculate the 16×16 block means to obtain a down-sampled version of an input frame. Then this small frame is filtered by a simple 3×3 Gaussian low-pass filter. Finally, we linearly interpolate it to form the frame with the original frame size. Fig. 3(a) shows a seriously blurred version of Foreman. The reason of removing high-frequency textures is to predict the effects of lossy compression as these parts are affected more. The other process is applying 16×16 block SVD after the block mean is removed. We then use the block mean and the important eigenvectors/eigenvalues to reconstruct the block. Such blocks will contain the significant content and can serve as reliable references to see what may be left after the coding. We simply include the first and second eigenvector pairs to reconstruct the block as shown in Fig. 3(b). Although the blocky artifacts can be found, the content can still be viewable. Given these two pre-processed or distorted frames, we calculate their quality degradation (D_{SSIM}), compared with the raw input frame. Then these two distortion measurements are combined to form the "content feature," which is used to evaluate the single parameter β in our model. Since it can be shown from Fig. 3 that the degrees of distortions in these two steps are quite different as the resizing results in a more serious quality degradation, the two evaluations are weighted and summed up to form our feature. In our training data evaluated in SSIM, the average distortion for resized frames, D^{resize}_{SSIM}, is around $K = 4$ times that of SVD processed frames, D^{svd}_{SSIM}. We thus calculate the "spatial feature", $F^{spatial}_{SSIM}$, by

$$F^{spatial}_{SSIM} = 0.2 \times D^{resize}_{SSIM} + 0.8 \times D^{svd}_{SSIM}, \quad (6)$$

which will be used to determine β.



(a) (b)

Fig. 3. Preprocessed Foreman by (a) resizing and (b) SVD.

Fig. 4 shows the relation between the extracted feature and β in the training data of I frames. We found that the data are clustered and can be fitted well by

$$\beta = 6.96 \times (F^{spatial}_{SSIM})^{0.68}. \quad (7)$$

Since only the intra-coding is applied in I frames, the feature for I frames, F^I_{SSIM}, is simply $F^{spatial}_{SSIM}$.

In P frames, the temporal information is required. As in the regular video coding, we apply the motion estimation with 16×16 blocks and the searching range set as 16×16 to form a motion compensated frame. Only the integer positions are searched. Like what we have done for I frames, the distortion of this compensated frame is computed to determine the temporal feature, $F^{temporal}_{SSIM}$. However, since the intra-coding may still be employed on P frames, we also calculate the spatial feature, $F^{spatial}_{SSIM}$, and use the average of the two features to determine (most of) the P frame feature by

$$F^P_{SSIM} = 0.5 \times F^{spatial}_{SSIM} + 0.5 \times F^{temporal}_{SSIM}. \quad (8)$$

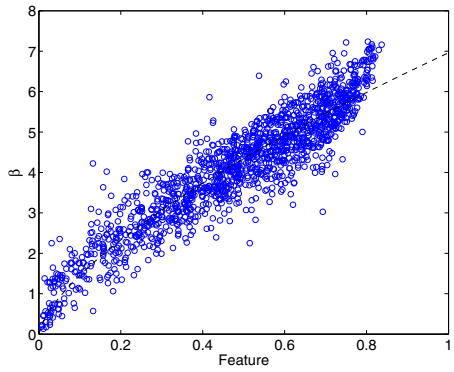


Fig. 4. The relationship between the extracted feature and β in I frames.

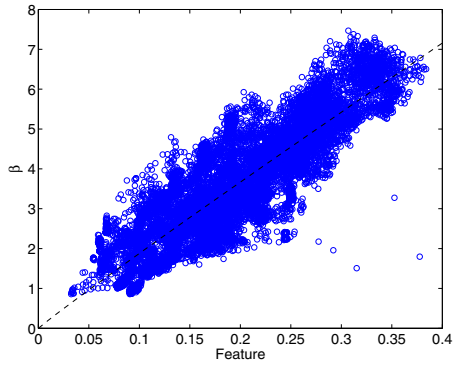


Fig. 5. The relationship between the extracted feature and β in P frames.

Fig. 5 shows the relationship between the feature and β in P frames. Although some outliers exist, the fitting can still be good enough to help choose suitable QP values of P frames. The fitting curve is

$$\beta = 17.32 \times (F^P_{SSIM})^{0.96}. \quad (9)$$

As mentioned before, we will calculate the feature to determine the single parameter β for each basic unit. Then the frame QP, QP_F, is determined such that the overall distortion will be as close to the target distortion as possible. That is

$$QP_F = \arg \min_{QP \in [0,51]} \sum_{i=1}^{12} (D^{(i)}(QP) - D^{target})^2, \quad (10)$$

where $D^{(i)}(QP)$ is the distortion of the i^{th} basic unit estimated by Eqs. (4) and (5). D^{target} is the target distortion. The use of 12 units helps to reduce the possible negative effect from the possible model inaccuracy of a single unit.

III. The Encoding Procedure

Our objective is to strictly maintain the quality of each frame. That is, after a target distortion is set, *e.g.* SSIM equal to 0.92, the distortion measurement of each decoded frame should reach the target value as close as possible so that the constant quality coding can be successfully achieved. With the proposed D-Q model, the selection of QP can be done in a straightforward manner. Given an input frame, the feature is computed to determine the D-Q relations of basic units and the frame QP will be chosen according to Eq. (10). There are a few issues that will determine the designs of our proposed encoding procedure. First, adjacent frames in a video usually have similar content, which will result in similar features. Then calculating features in each frame doesn't seem that necessary. The spatial feature $F^{spatial}$ is relatively efficient but the temporal feature $F^{temporal}$ is more time-consuming because of the motion estimation. Therefore, if we can reuse the feature of a previous frame with the similar content for computing the model parameter β, the whole encoding procedure will be more efficient. In other words, the spatial and temporal features of a frame will only be re-calculated if such feature with the same frame type or similar content is not available before. Second, the quality of the referenced frame will affect that of the currently encoded frame. Especially when a scene change frame appears and its QP is not appropriately assigned, the quality of subsequent frames may be poor. Large quality variations may also be observed. Our strategy is to apply the scene change detection to determine the so-called key frames to build the D-Q model. We will then encode these frames carefully, probably with two runs, so that the quality of subsequent frames can also be maintained. Third, the coding performance of the previous frame of the same type will be referenced for the model adjustment. As mentioned before, the content of adjacent frame will be similar. If the frame coding types are the same, the coding results of the previous frame can serve as a good indication of the model accuracy. We may make use of this information to adjust our model so that the single-run coding may work as well as the multiple-run coding. The encoding procedure will be detailed as follows.

For the scene change detection, we apply a simple process by examining the luminance histograms of adjacent frames. The Bhattacharyya distance of two histograms is calculated and compared with a threshold. If the difference is larger than the threshold, a scene change is detected and the key frame is extracted. It should be noted that, although the key frame may need to be encoded as a P frame, we only use the spatial feature $F^{spatial}$ to calculate β, instead of using the P frame feature F^P in Eq. (8), because a large number of intra-coded blocks will appear in this frame. After using $F^{spatial}$ to determine the D-Q relation and the frame QP for encoding this frame, we usually encode this frame once again if the resulting quality of this decoded frame is not satisfactory. This two-pass encoding is to ensure that these important scene-change frames have the target quality after the encoding. The model will be slightly adjusted according to the first-run encoding result. We call this process as the model update, which is actually adding an adjusting factor θ defined by

$$\theta = \frac{D^p(QP_F)}{e^{(a*\beta+b)} \times QP_F^{\beta}}, \quad (11)$$

where a and b are the trained variables listed in Eqs. (4) and (5) so the denominator is the predicted distortion by our model and $D^p(QP_F)$ is the resulting distortion by using QP_F to

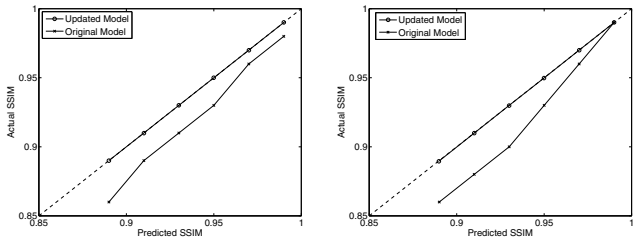


Fig. 6. The comparison of coding results of Foreman by using the original model and updated model in the cases of (a) SSIM in I frames and (b) SSIM in P frames.

encode the frame in the first run. In the second-run encoding of this frame, the model becomes

$$\theta \times D_{SSIM}^{I/P}, \tag{12}$$

where D_{SSIM} is defined in Eqs. (4) and (5), and a better QP_F can then be chosen accordingly. In other words, we simply adjust the parameter α in Eq. (2) and this strategy is quite effective. Fig. 6 shows the comparison of coding results on Foreman by using the original model and updated model with θ. In Figs. 6(a), we encode all the frames by using the intra coding only. In Fig. 6(b), only the first frame is an I frame and the other frames are coded as P frames. The qualities of P frames are then averaged. We select Foreman in this test since it contains large content variations and our original model doesn't perform that well. By using the simple scaling factor θ, the predicted quality, measured in SSIM, will be close to the actual quality after the model adjustment of the second-run encoding in some of the key frames.

For other frames, we will use the coding result of the previous frame with the same frame type as the reference to adjust our D-Q model. That is, θ will be computed by dividing the resulting distortion of the previous frame (*i.e.* $D^p(QP_F)$ in Eq. (11)) by the predicted distortion so that most of the frames will be encoded only once. As mentioned before, only the spatial feature $F^{spatial}$ will be used to find β in the scene-change frames. There will be a couple of special cases for other frames. 1) For the first P frame after the scene-change frame, since its temporal feature $F^{temporal}$ is not available, we will calculate its own feature F^P. In addition, since the previous P frames do not have the similar content, this P frame may be encoded twice without referring to the coding results of previous frames. 2) For the first I frame after the scene-change frame, we will calculate its own $F^{spatial}$ to calculate β and may also encode this frame twice to use its own first-run coding results for the model adjustment. To be more specific, the features will be computed and the coding may be applied twice in the following three cases: 1) The scene-change or key frame, 2) the first I frame after the key frame and 3) the first P frame after the key frame. For most of the other I/P frames, we basically employ the existing features and use the coding results of the frames with the similar content and with the same frame type for the model adjustment. The calculation of the features will not be applied repeatedly. Finally, to achieve the very consistent video quality, the coding result of each frame will be checked. If the result deviates from the target too far, we may encode that frame once more and again the model is adjusted by Eq. (12). It should be noted that no more than two encodings will be applied on a frame to maintain the efficiency of the proposed method. The flowchart of the encoding process is demonstrated in Fig. 7.

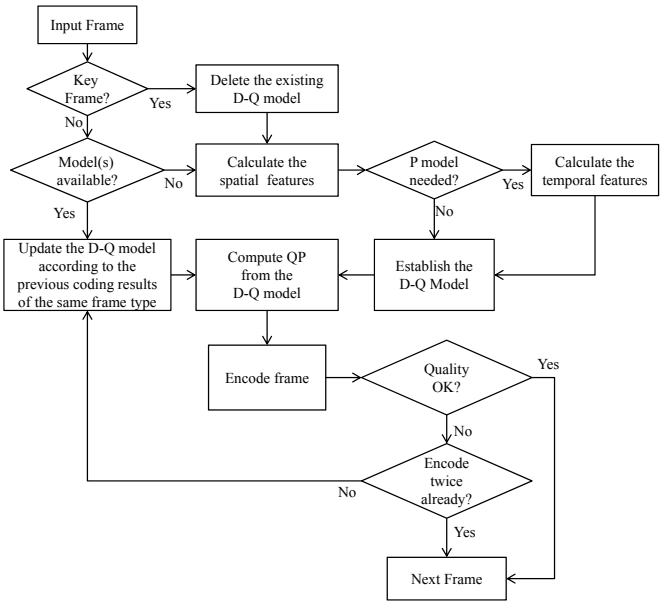


Fig. 7. The flowchart of the encoding procedure.

IV. Experimental Results

We have implemented our scheme on the JM 17.2 reference software of H.264/AVC to evaluate the performances of our proposed D-Q model and encoding procedure. The settings of JM 17.2 are as follows:

1) Baseline profile with I/P frames.
2) Rate distortion optimization is enabled.
3) The motion Search range is ± 16.
4) Fast Full Search algorithm is used.
5) The number of the reference frame is 1.
6) CAVLC coding is enabled.
7) De-blocking filter is used.

We set the target SSIM as 0.91, 0.95 and 0.99 to test the feasibility of our scheme on different quality measures. SSIM is calculated in 8×8 blocks without overlapping. Eight CIF videos including Coastguard, Monitor, Table, Foreman, Mobile, Stefan, News and Paris, each with 300 frames, are used in our experiments. Fig. 8 shows the performance of constant quality video coding measured in SSIM. We can see that the resulting quality can achieve the target quality in all the cases. When the target quality is set lower, the variation of SSIM becomes larger because of the wider range of QP. The variations are more obvious in the latter part of Foreman because of the fast camera motions. The two-pass encoding

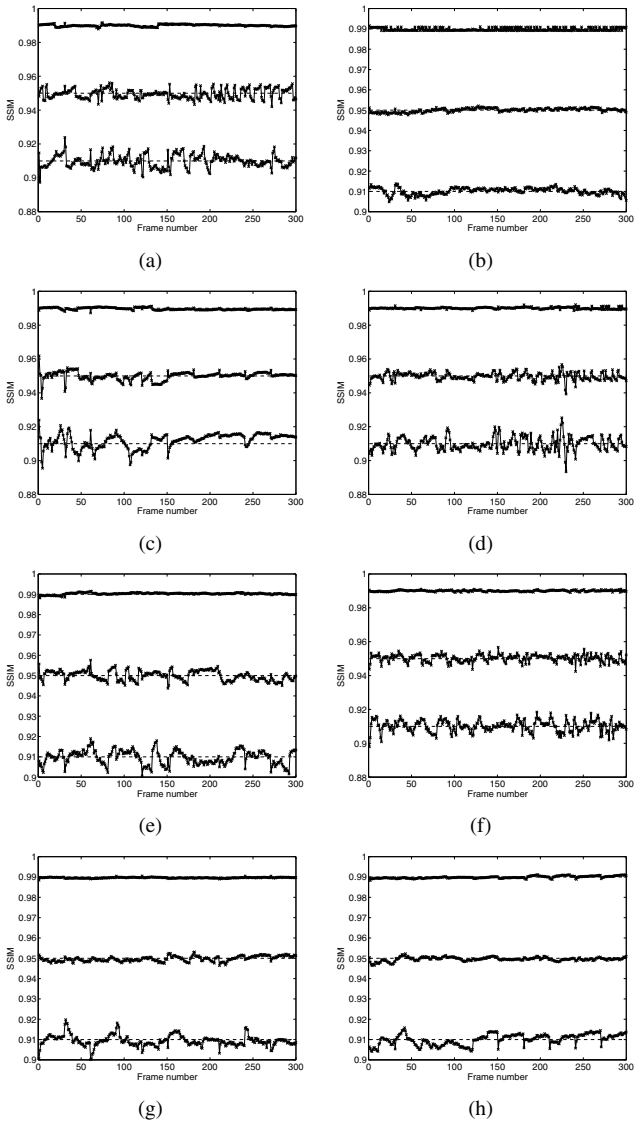


Fig. 8. The performances of constant quality (SSIM) video coding of (a) Coastguard, (b) Monitor, (c) Table, (d) Foreman, (e) Mobile, (f) Stefan, (g) News and (h) Paris.

is not applied very often and the most frequent case is when the target SSIM is set as 0.91 in Foreman, in which only 16 out of 300 frames are encoded twice. In other videos such as Monitors and Mobile, except the first two frames, which are the first I and P frames and do not have any previous results for referencing, other frames are encoded just once.

V. Conclusion

In this research, a frame quality control mechanism for H.264/AVC is proposed. A suitable Quantization Parameter (QP) can be assigned in each frame so that the target frame quality can be achieved. A single-parameter D-Q model is derived and the model parameter can be determined by the frame content. The results by using such quality measurement as SSIM verifies the feasibility of our proposed method. We will extend to test some more quality metrics to further prove the generality of this frame work.

Acknowledgment

This research was supported by the National Science Council in Taiwan, under Grant NSC100-2221-E-008-120-

References

[1] T. Wiegand, G. 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, Sep. 2003.

[2] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. on Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[3] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. on Broadcasting*, vol. 50, no. 3, pp. 312–322, Sep. 2004.

[4] T. J. Liu, K. H. Liu, and H. H. Liu, "Temporal information assisted video quality metric for multimedia," in *IEEE International Conference on Multimedia and Expo (ICME)*, Jul. 2010, pp. 697–701.

[5] A. K. Moorthy and A. C. Bovik, "Efficient video quality assessment along temporal trajectories," *IEEE Trans. Circuits and systems for video technology*, vol. 20, no. 11, Nov. 2010.

[6] Y. H. Huang, T. S. Ou, P. Y. Su, and H. H. Chen, "Perceptual rate-distortion optimization using structural similarity index as quality metric," *IEEE Trans. Circuits and systems for video technology*, vol. 20, no. 11, Nov. 2010.

[7] T. S. Ou, Y. H. Huang, and H. H. Chen, "SSIM-based perceptual rate control for video coding," *IEEE Trans. Circuits and systems for video technology*, vol. 21, no. 5, Mar. 2011.

[8] K. L. Huang and H. M. Hang, "Consistent picture quality control strategy for dependent video coding," *IEEE Trans. Circuits and systems for video technology*, vol. 18, no. 5, Mar. 2009.

[9] S. Ma, W. Gao, and Y. Lu, "Rate-distortion analysis for H.264/AVC video coding and its application to rate control," *IEEE Trans. Circuits and systems for video technology*, vol. 15, no. 12, pp. 1533–1544, Dec. 2005.

[10] N. Kamaci, Y. Altunbasak, and R. M. Mersereau, "Frame bit allocation for the H.264/AVC video coder via cauchy density-based rate and distortion models," *IEEE Trans. Circuits and systems for video technology*, vol. 15, no. 8, pp. 994–1006, Aug. 2005.

[11] C. Y. Wu and P. C. Su, "A content-adaptive distortion-quantization model for intra coding in H.264/AVC," in *The 20th IEEE International Conference on Computer Communication Networks (ICCCN 2011)*, Maui, Hawaii, Jul. 2011.

[12] F. D. Vito and J. C. D. Martin, "PSNR control for GOP-level constant quality in H.264 video coding," in *IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, Athens, Greece, Dec. 2005.

[13] B. Han and B. Zhou, "VBR rate control for perceptually consistent video quality," *IEEE Trans. Consumer Electronics*, vol. 54, no. 4, Nov. 2008.