SUBJECTIVE QUALITY EVALUATION OF H.264 AND H.265 ENCODED VIDEO SEQUENCES STREAMED OVER THE NETWORK

Dipendra J. Mandal and Subodh Ghimire

Department of Electrical & Electronics Engineering, Kathmandu University, Dhulikhel, Nepal

ABSTRACT

Streaming of high-quality video contents as part of multimedia communication, has become essential nowadays. Video delivered over the network suffers from different kind of impairments, which degrades its quality. Such network impairments effect differs among different types of codec used. This paper present the effects of network degradation factors such as packet loss and jitter over H.264 and H.265 encoded video sequences. In addition to the codec used, we also focused on the different level of temporal and spatial aspect within the videos. Among different basic test methods, double stimulus impairments scale was used to complete the experiment as subjective measures of assessment metric from user's perspective.The result illustrates that differently encoded video sequences react differently to the network impairments and are very sensitive to a transmission error. Similarly, it also shows that user's experience is affected according to the motion level of video.

KEYWORDS

Video Quality, H.264, H.265, Subjective Evaluation, Packet loss, and Jitter.

1. INTRODUCTION

Due to rapid growth in Internet use and easy access to more users, the deployment of the digitally compressed video is also rapidly increasing. Distribution of Video content is becoming one of the most important applications for the real-time system. Visual quality assessment has been so far one of the most intriguing challenges in the media environment. Quality can be defined in a number of ways, depending on the application of the multimedia service and the end-user of the audio-visual content. Network impairments such as packet loss, packet delay, and jitter play a dynamic role to degrade the quality of received video when transmitted over the network. Because of variation in coding techniques (such as predictive coding, variable length coding) used by encoders, different compressed video streams can have different sensitive nature to the transmission error. In few cases single corrupted bit within the video stream can lose the synchronization so that even successive correctly received bit becomes useless and result in a drastic reduction in the quality of the video. Due to which, quality of the video is the major issue. Digital video quality measurements must be based on the perceived quality of the actual video being received by the user of the digital video system because the impression of the end user is nevertheless the important factor [1].

A major challenge for digital video is that raw or uncompressed video requires lots of data to be stored or transmitted due to which, compression is needed. Compression aims at lowering the total number of parameters required to represent the signal while maintaining good quality. Currently, several video compression standards exist for different multimedia applications. Each standard may be used in a range of applications but is optimized for a limited range. H.261,

H.263, H.263 designed by ITU (International telecommunication Union) is aimed at low-bit-rate video applications. MPEG standards are defined by ISO (International Organization for Standardization). MPEG-2 is aimed at high bit rate high-quality applications, and MPEG-4 is aimed at multimedia applications including streaming video applications on mobile devices [2]. H.264 also referred to as MPEG-4 Part 10 Advanced Video Coding appeared in the mid-2000s. It has greatly improved the coding performance over MPEG-2 and MPEG-4 Part 2 [2][3], H.264 offers the same compression capability at approximately half the bitrate. Video can be encoded using the different profiles of H.264 with the baseline profile, the file can be easily played even on devices with low computational power whereas High and Main profiles target the high computational devices because of its high compression capabilities. Now, a new standard H.265 also called High-Efficiency Video Coding (HEVC) has been developed that promises a further factor of two improvements in compression efficiency compared with the previous standards[5][6].

When network congestion occurs it results to packet drops. The effects of packet loss on real-time multimedia application are critical and the effect of extensive packet loss on video is acute. If packet loss happens, some part of the video cannot be decoded and displayed. Similarly, packet delay variation also known as jitter is variation between packet arrival times at receiving end, this occurs due to the variability in queuing and propagation delays. A small value of jitter can be tolerable but increased value causes decoding error and causes quality degradation [7]. Therefore, it is important to understand the effect of packet loss and jitter on the perceived quality of video applications.

In this paper, we have presented the details and results of the subjective quality evaluation performed at the laboratory of Kathmandu University. The aim of this paper is to evaluate the quality of various encoded video sequence such as H.264 with two different profiles called Baseline and High and H.265 when streamed over the network. Network effects such as various levels of packet drop and jitter over the encoded video were studied and compared with the objective analysis previously performed in our previous work [8].

The rest of the paper is organized as follows: Section II presents the description of various types of video quality assessments. In section III, we have described the detail of video specification along with the subjective methods used, laboratory setup and statistical data processing approach. The result from the experiment with its explanation is presented in section IV. At the end of this paper, concluding remark is given.

2. VIDEO QUALITY ASSESSMENT

In this section, we have described the theory of different video quality evaluation methodology.

There are two standardized test methods to measure the quality of multimedia experience.

1. *Objective Quality Matrix*
2. *Subjective Evaluation*

Objective video quality evaluation techniques are mathematical models that approximate the results of the subjective quality assessment. A quality matrix such as peak signal to noise ratio (PSNR), Mean square Error (MSE) and structural similarity (SSIM) are used. These matrices are evaluated using the computer program and hence is fast and easy to conduct the experiment. In contrast, subjective video quality experiments involve the human subjects to watch the video sequence and to rate their quality in numerical values instead of computers to measure the quality level and results are expressed by the means of mean opinion score (MOS). MOS is an average of

scores provided by the number of users for the same set of experiment. Different types of subjective methods are explained in the ITU-R. BT. 500 [9] and ITU-T Rec. P.910 [10]. There are numerous studies performed in a controlled research laboratory setting to evaluate how quality affects the user experience for different types of media contents. Few of them are described in [11] [12].

There are wide varieties of basic test methods that have been used for the subjective quality analysis. Some examples are:

- *Double Stimulus Continuous Quality Scale (DSCQS):* Reference and test video sequences are shown twice to the observer and are asked to rate the quality.
- *Double Stimulus Impairment Scale (DSIS):* Reference and test video sequence are shown only once and are asked to grade the video. This type of methods is well suited for clearly visible impairments such as artifacts caused by transmission errors.
- *Single Stimulus Continuous Quality Evaluation:* In this method instead of showing the separate short video sequence pair, observers are shown a video session of typically 20-30 minutes also reference video is not shown.
- *Absolute Category Rating:* This is the single- stimulus method. Observers rate each test video sequences individually without comparison to an explicit reference. Since no reference is shown, from a time perspective it is considered as efficient method over DSIS or DSCQS.
- *Pair Comparison:* In this, test videos taken from the same source but under different conditions are paired in many possible combinations and observers are asked to make a preference judgment for each pair.

All above test methods have different rating scale which includes both continuous and discrete level and are used for different application [13] and a minimum of 15 observers are recommended in order to conduct the experiment. Details about rating scale and analysis of above methods are discussed in [14].

3. EXPERIMENT DESIGN AND EVALUATION

In this section details of video sequences used for the test is explained. We have further introduced the design and facts of subjective methods and statistically data processing methods used and also all the specifics related to the laboratory setup has been discussed.

3.1. VIDEO SEQUENCES

We have used the two videos for this experiment. Both the video sequence has encoded using the H.264 baseline, H.264 High profile and H.265 and has a different level of motion. These video clips are reconstructed and captured after being streamed over the network having a different level of packet drops and packet variation. Details of the video network emulation setup and procedure used to obtain video sequences can be seen in [8].Table 1 shows the details of encoded parameters for video samples.

Table 1. Video encoding parameters

Video Files / Parameters	Video_1	Video_2	Video_3
Codec	H.264	H.264	H.265
Profile	Baseline	High	-
No. of frame	270	270	270
Frame Resolution	1280*720	1280*720	1280*720
Frame Rate	29.97	29.97	29.97
Chroma Format	4:2:0	4:2:0	4:2:0
No. of I, P & B frames	I=18, P=252	I=2, P=268	I=2, P=40, B=228
Duration	9.74 sec	9.74 sec	9.74 sec
File Size	25.0 MB	1.26 MB	196 KB

3.2. SUBJECTIVE QUALITY EVALUATION

The method we selected for our experiment is DSIS. In this method, two videos were shown to the observer, one the unimpaired (reference video) and the impaired (test) video. The reference video was made known to the observer and shown before the test video. The general arrangement of a test system for this method is shown in figure 1. After viewing the both video sequence, the observer was asked to grade the test video compared to the reference. The grade scale used is given in table 2 and presentation of video sequence used in this experiment is shown in figure 2.

Table 2. Grade scale with description for DSIS methods

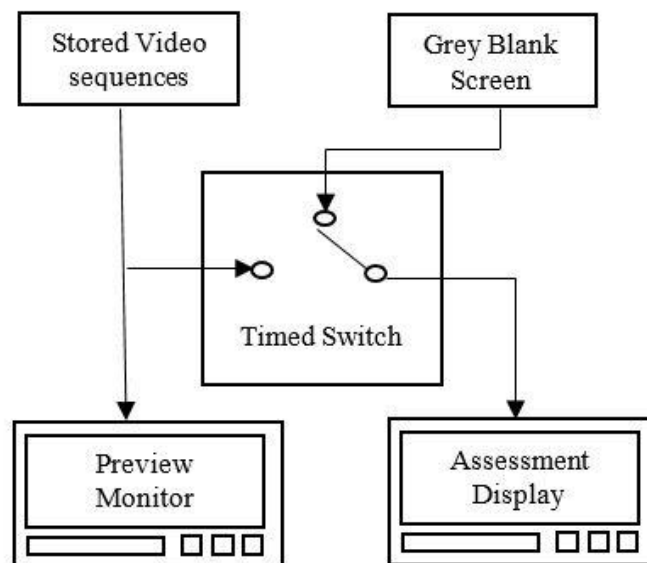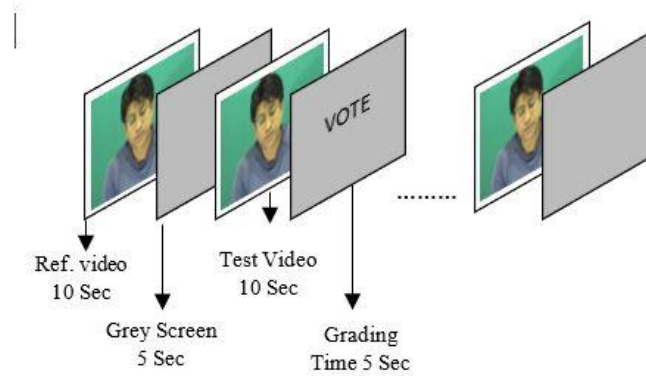| Scale	Description
5	Imperceptible
4	Perceptible but not annoying to observer
3	Slightly annoying to observer
2	Annoying to observer
1	Very annoying

Figure 1. Arrangement for test system for DSIS method

Figure 2. Presentation sequence with duration used in experiment

3.3. LABORATORY EXPERIMENTAL SETUP

The test has been carried out according to the recommended guidelines of BT.500-11 [9], which suggest standard viewing conditions, criteria for the observer and test material selection including experimental procedure and data analysis methods. Observer was seated directly in line with the center of the video display monitor at a specified distance, which was 7-8H where H is the height of the LCD screen. Details of LCD display device used during the experiment is provided in table 3. The sets of the video were evaluated in an experiment performed at the laboratory of Kathmandu University, Nepal. Observer taking part in the experiment were from the Department of Electrical and Electronics mostly students of undergraduates and faculty. The subject pool consisted both male and female and each was tested for color blindness, this was performed using a Snellen chart [15], color blindness using Ishihara plates [16]. Each observer was briefed about the test prior to the experiment with short training session of 6-10 minutes. This was done in order to stabilize the observers and data issued from this session was not included. Training session includes an explanation about what the observer are going to see, what they have to evaluate and how to evaluate the video with a different level of distortion. Distorted video sequences were shown in random order. The experiment was running over 5 days with 6-8 observers per day this was done to avoid the viewer's exhausting which may affect their evaluation ability. Figure 3 shows the picture of laboratory setup for performing the subjective test experiment.

Table 3. Testbeds screen details

Size	18.5 Inches
Resolution	1366*768
Contrast Ratio	5000:1
Screen Mode	WXGA
Aspect Ratio	16:9
Brightness	$250/m^2$

Figure 3. Subjective test laboratory setup picture

3.4. STATISTICAL DATA PROCESSING

In our subjective lab experiment, altogether 34 observers participated, they rated 62 video sequences corresponding to the two different video contents reconstructed after a different level of network impairments. Out of 34 observers one was found to be color blind and reading from him was not taken for the calculation. Data obtained from the observer was obtained was an integer value between 1 to 5 and need to be summarized using statistical techniques. For this first normality test was performed to verify the obtained data and this was done using the β2 test (Kurtosis coefficient of the function) described in ITU-R BT.500-11. If the value falls between 2 and 4 then the distribution is normal. Scores provided by the observer was found to be close to the normal distribution. Secondly, the outlier detection was performed for each observer. This is described in Annex 2 of ITU-R BT.500-11. If the score of observer deviates strongly from the average score for given presentation, then the observer was discarded. A maximum of 3 observers was discarded after performing this analysis. Finally, after the filtration of raw data, mean opinion score (MOS) for each video sequence was evaluated, which is mathematically expressed in equation 1.

$$MOS_j = \frac{\sum_{i=1}^{N} O_{ij}}{N}$$

(1)

Where,

N: is the total no. of the valid observer after discarding the outlier.
Observer = i= 1, 2 ...... N
Test Sequence = j= 1, 2 ...... N
O_{ij}: is the score provided by observer i for test sequence j

Finally, MOS for each video sequences were plotted using the confidence interval (CI) which is derived from the standard deviation and size of the sample. For the normally distributed data with sample size N, MOS_j and standard deviation S_j, CI is defined as $(MOS_j \pm \delta_j)$, where δ_j is expressed in equation 2.

$$\delta_j = z \frac{S_j}{\sqrt{N}}$$

(2)

For the subjective quality data analysis, it is recommended to calculate the 95 % CI for each point of the test. For 95 % CI with a normal distribution, the value of z in equation 2 is 1.96.

4. RESULTS AND DISCUSSION

In the initial phase of our experiment, we studied the effect of packet drop over MOS for video with a different level of motion streamed over the network. Figure 4 shows the effect of packet drop over MOS for both high and low level of temporal variation in the video. The outcome shows that video with low temporal variation has better MOS over high temporal variation video.The plot also shows the CI result, as it can be seen from small confidence interval, the result obtained from different subjects are reliable and also variation between the score obtained from various subjects are very small.
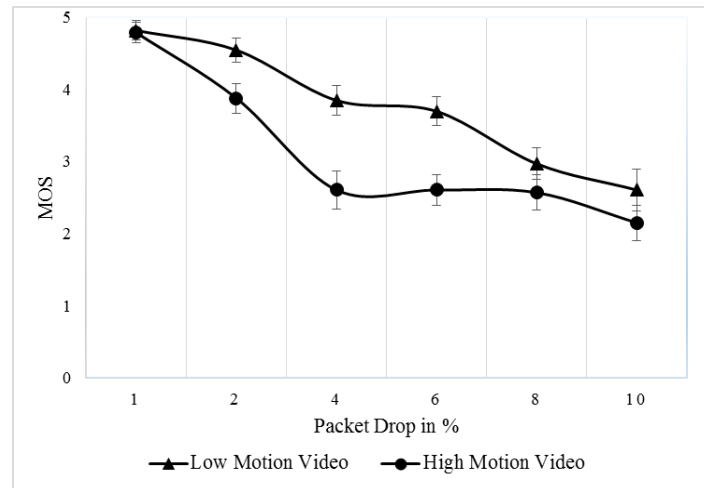

Figure 4. Packet drop over MOS for video with low and high level of motion (95 % CI)

We also studied the effect of packet drop over user opinion for H.264 with Baseline profile, H.264 with High profile and H.265 encoded video sequence. Figure 5 and figure 6 demonstrate the result for both low and high motion video respectively. It shows that codec with high compression efficiency i.e. H.265 is affected more severely for given packet drop as compared to H.264 with a different profile. The main reason behind this is due to the difference in a group of picture structure for a different encoded video sequence.
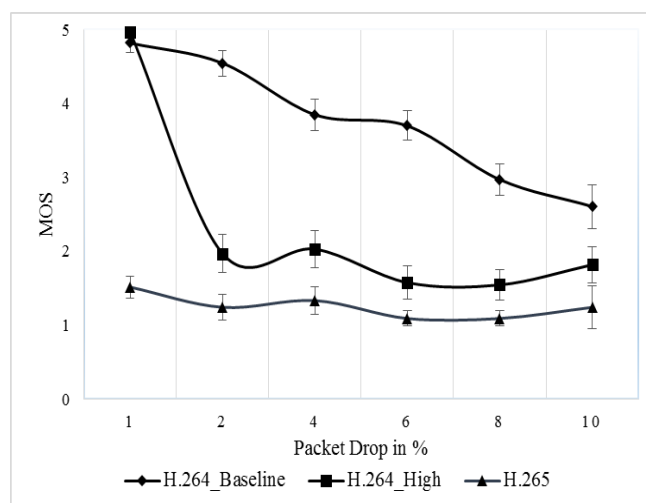

Figure 5. MOS Vs. Packet drop for three different encoded video sequences (Low motion)
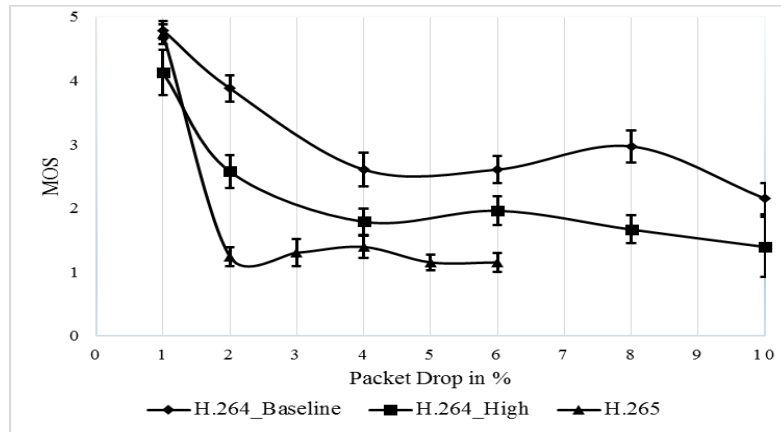
Figure 6. MOS Vs. Packet drop for three different encoded video sequences (High motion)

To analyze the effect of jitter on the video quality from a user perspective, few tests were made in which observer rated the video sequence with different level of jitter variation i.e. we repeated the measurement for 1, 5, 10, 20, 50, 100, 200, 500 and 1000 millisecond delay variation. The result shows that video with the high temporal aspect is more sensitive to increased jitter level compared to low temporal aspect video sequences. This is lustrated in figure 7 with 95% CI, which shows that the result obtained from different subjects are reliable.

The experiment also shows that effect of jitter for video sequence encoded with different video codec when transmitted over the network. Graphical representation plotted from the data obtained is shown in figure 8. The result shows that the video encoded with baseline profile of H.264 is more prone to the added jitter compared to others. Result also shows that jitter has an almost negligible effect on all three encoded video sequences up to 10 ms of added jitter to the network after that H.264 encoded video with profile baseline is affected drastically in compare to other two. Also in presence of high jitter even other two encoded video sequences are affected equally.
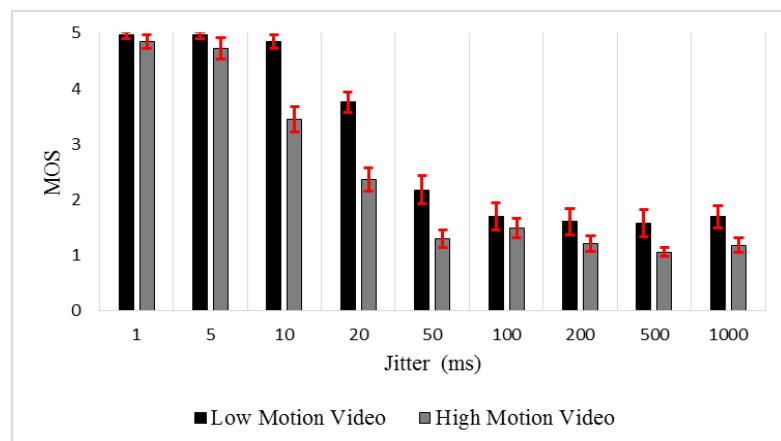

Figure 7. Result of Jitter over MOS for video with two different level of motion
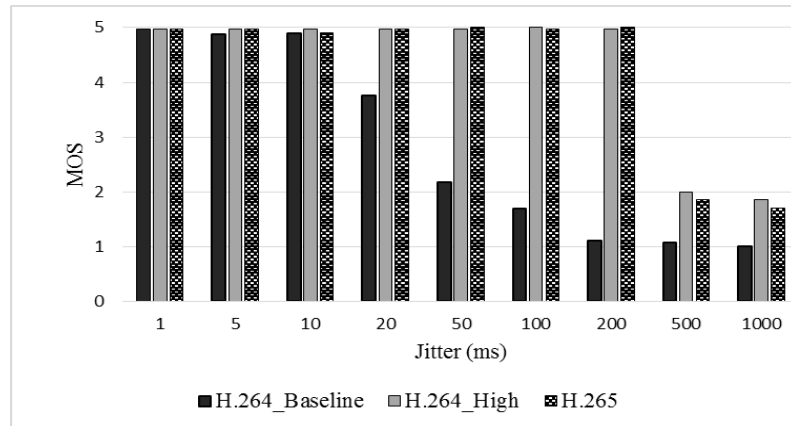
Figure 8. Effect of jitter over MOS for three different video codec

5. CONCLUSION

In this paper, we presented a detailed description of subjective video quality assessment performed for three different encoded video streams when streamed over the error-prone network. The results obtained is close to testing output obtained in our previous work using objective evaluation. The obtained result shows that video encoded with H.265 is affected more with packet drop compared with its predecessors. However, in presence of network jitter, H.264 encoded videos suffers more. This clearly indicates that although newly developed codec H.265 is twice bandwidth efficient than H.264 but is more sensitive to the network error and suffers more severely when a small amount of error is introduced into the network.

REFERENCES

[1] J. Klaue, B. Rathke and A. Wolisz, "Evalvid -A Framework for Video Transmission and Quality Evaluation," in 13th International Conference on Modeling Techniques and Tools for Computer Performance Evaluation, September 2003.

[2] I. Ahmad, S. Wei, Y. Sun and Y. Zhang, "Video Transcoding: An Overview of various Techniques and Research Issues," IEEE Transactions on Multimedia, vol. 7, no. 5, October 2005.

[3] K. R. Rao, D. N. Kim and J. J. Hwang, Video coding standards: AVS China, H.264/MPEG-4 PART 10, HEVC, VP6, DIRAC and VC-1, Illustrated ed., Springer Science & Business Media, 2013, pp. 271-272.

[4] "ISO/IEC 14496-10 and ITU-T Rec. H.264 advance video coding," 2003.

[5] D. Hingole, "H.265 (HEVC) Bitstream to H.264 (MPEG 4 AVC) Bitstream Transcoder," The University of Texas, Arlington, December 2015.

[6] "HEVC Demystified: A Primer on the H.265 Video Codec," Elecmental technologies, Inc., 2013.

[7] B. Tulu and S. Chatterjee, "Internet-based telemedicine: An empirical investigation of objective and subjective video quality", Decision Support Systems. 2008, vol. 45, pp. 681-696.

[8] D. J. Mandal and S. Ghimire, "Effect of Packet Drop and Jitter on Perceived Video Quality for various Encoded Video over Streaming Network", August 16 Volume 4 Issue 9, International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC), ISSN: 2321-8169, PP: 07-11

[9] "13. ITU-R Recommendation BT.500-11. Methodology for the subjective assessment of the quality of television pictures," International Telecommunication Union , Geneva, 2002.

[10] "14. ITU-R Recommendation P.910. Subjective video quality assessment methods for multimedia applications," International Telecommunication Union , Geneva, 2008.

[11] K. yang, C. Guest, K. EI-Maleh and P. Das, "Perceptual Temporal Quality Metric for Compressed Video," in IEEE Transaction on Multimedia , November 2007.

[12] I. Ceaparu, J. Lazar, K. Bessiere, J. Robinson and B. Shneiderman, "Determining Causes and Severity of End-User Frustation," International Journal of Human-Computer Interaction, 2004.

[13] S. Winkler, "Digital Video Quality – Vision Models and Metrics", John Wiley & Sons, 2005

[14] S. Winkler, "On the properties of subjective rating in video quality experiments," in International Workshop on Quality of Multimedia Experience (QoMEX), San Diego, July 2009.

[15] C. Boyer, "Pro Visu," HON Foundation, [Online]. Available: https://www.provisu.ch/en/screening/tests.html. [Accessed August 2016].

[16] "Color Blindness- Ishihara Color Test," [Online]. Available: http://www.colour-blindness.com/colour-blindness-tests/ishihara-colour-test-plates/. [Accessed August 2016].

AUTHORS

Dipendra Jee Mandal received his bachelor in Electrical and Electronics Engineering with specializing in Communication from Kathmandu University, Dhulikhel, Nepal in 2010. He is currently enrolled as MS by Research student at same university. He worked as researcher at Department of Electrical and Electronics Engineering, Kathmandu University from 2010 to 2013. His current research interest is video coding, quality assessment and digital image processing.

Subodh Ghimire received his bachelor in Electrical and Electronics Engineering with specialize in Communication from Kathmandu University(KU), Dhulikhel, Nepal in 2001 and MS in Electrical in 2003 from San Jose State University, CA, USA. He is currently working as Assistant Professor at Department of Electrical and Electronics Engineering, KU.

The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper.