A PROPOSED FRAMEWORK FOR VIDEO QUALITY ASSESSMENT USING WAVELET AND STATISTICAL MEASURES

Shadiya P
Dept. of Electronics and Communication Engineering,
MES College of Engineering,
Affiliated to Calicut University,
Kuttipuram, kerala, India
shadiyasmail@gmail.com

Balachandran K P
Dept. of Computer Applications,
MES College of Engineering,
Affiliated to Calicut University,
Kuttipuram, kerala,India
kpbala@gmail.com

Abstract— Fast growing video technologies make the world closer to digital videos. There are many video applications of them some are mobile video streaming, video conferencing. Video quality evaluation is very important in the design and optimization of wireless video processing and transmission systems. To evaluate the quality of video, there are various subjective methods and objective video quality assessment (VQA) algorithms that have been developed with varying computational complexity and accuracy. These VQA algorithms were tested for their correlation with human perception. In this paper, introducing a Full Reference (FR) algorithm for assessing the quality of a video that gives better results with high accuracy and low complexity. Due to these features it can be used for real time applications.

Keywords- Full-Reference; Video Quality Assessment; Peak Signal to Noise Ratio; Structural Similarity Index Measure

I. INTRODUCTION

Video transmission is one of the most important application of telecommunication systems and are supporting different kinds of real-time transmissions.Video Streaming is becoming prominent in current generation mobile wireless network. Even with the current high-end video technology, streaming are limited to low-quality video due to the bandwidth availability to the end user.Video compression methods are used to deliver a fair quality of video over this modern telecommunication system.

A digital video passes through numerous processing stages before it finally reaches the end-user. The original video sequence at the transmitter end is passed through an encoder which compresses and restructures the video sequence, which is then passed over a channel. At the receiver end, a decoder decompresses the sequence into a format visible to the end user. Throughout this process distortions are introduced in the video stream which can produce visually annoying artifacts at the end-user. The encoder, the channel, the decoder, and the display can introduce distortions in the video sequence. Encoder errors may include blocking artifacts, blurring, discrete cosine transform, basis image effect, color bleeding, ringing, and so on due to restrictions on bit-rate and errors in the motion estimation process. The channel, being inherently noisy, can corrupt the video in many ways.

The end user of the videos are human observers, hence human opinion is ultimate. One method to assess the quality of a video is collecting the human opinion. This is called subjective method. But this is challenging because it require trained experts to judge it and it is not easy to implement, since it is time consuming and difficult. Another method to assess the quality of video is objective method. Objective methods are mathematical models based on certain criteria's and metrics, which can evaluate the quality of a video objectively and automatically by a computer program and it approximate the result of subjective quality assessment. The performance of an objective video quality metric is evaluated by computing the correlation between objective scores and subjective test results. Correlation coefficients are used to find the correlation of objective algorithms with human opinions.Video Quality Experts Groups (VQEG) is the principal forum that validates objective video quality metric models that result in International Telecommunication Union (ITU)

recommendations and standards for objective quality models for both television and multimedia applications [1],[2].

II. PREVIOUS WORKS

Video Quality Measurement metrics can be categorized as full-reference (FR), reduced-reference (RR), and no-reference (NR) based on the availability of original video. In Full Reference, Evaluation is done by comparing a degraded video with the reference video. This is highly accurate objective assessment method. But it requires a very large amount of data from the original video and is mainly used for Codec optimization, Off-service quality check, Content-encoding quality monitoring at head end and On-site quality check. In Reduced Reference, evaluation is done by comparing processed video subjected to distortion by coding and transmission losses with a small amount of information extracted from the source video. This is generally specific to an application. It provides video quality comparison at network nodes. But it is not as accurate as the FR model because the model requires only a small amount of feature data from the source video. In this a mean of transmitting the feature data is required. This is mainly used in In-service quality monitoring at user end and Content-encoding quality monitoring at head end. In No Reference, design of algorithms is extremely challenging and little progress has been made. This evaluates video quality on the basis of processed frames without any original information and it can be applied in a great many environments. But it is less accurate in evaluating the quality of video than FR and RR. In-service quality monitoring at user end is one of the application [1].

Objective methods can also be classified in terms of their usability in the context of adaptive streaming solutions as out-of-service methods and in-service methods: In out of- service methods there are no time constraints and original sequence are required. Full-reference visual quality assessment metrics and high-complexity non real-time RR and NR metrics fall within this class. In in-service methods this method place strict time constraints on the quality assessment and is performed during streaming applications [2].

Peak Signal to Noise Ratio (PSNR) is a traditional point based metrics used commonly as a measure of quality in video processing and this is computationally simple. The signal in this case is the original data, and the noise is the error introduced by compression. It is a measure of the mean square- error between the two signals being compared. For video-sequences, the PSNR is calculated for each frame and then averaged across frames. The main disadvantage is that it can't correlate well with the visual perception [2].

There are some metrics based on Natural Visual Statistics, which uses statistical measures such as mean, variance, covariance and distributions. Single-Scale Structural Similarity Index (SS-SSIM) is designed for still images, based on the principle that HVS is highly adapted for extracting structural information. It is defined as a product of a structure term, an intensity term, and a contrast term. SSIM Index was specific to still image quality assessment. The quality of the image is defined as the average of the quality map, i.e. the mean SSIM (MSSIM) index [2].

For video sequences the Video Structural Similarity Index Measure (VSSIM) metric measures the quality of the distorted video in three levels namely the local region level, the frame level, and the sequence level. In the local region level, local sampling areas are extracted and calculate SSIM. The local quality index is obtained as a function of the SSIM indices for the Y, Cb, and Cr components. At the second level, the local level quality values are weighted to give a frame level quality measure. Frame level quality measures are in turn weighted to obtain the overall quality of the video sequence [3].

Multi Scale Structural Similarity Index Measure (MS-SSIM) provides more flexibility by incorporating the variations of the image resolution and viewing conditions, which is an extension of the single-scale approach used in SSIM and it performs better relative to human opinion than the SS-SSIM index on images. MS-SSIM method applies a low pass filter to the reference and distorted images and down samples the filtered images by a factor of two. At the m^{th} scale, contrast and structure comparisons are evaluated. The luminance comparison is computed at scale M (i.e. the highest scale obtained after M – 1 iterations). This metric outperform the SSIM index and many other still image quality assessment algorithms. The MS-SSIM index can be extended to video by applying it frame by frame on the luminance component of the video and the overall MS-SSIM index for the video is computed as the average of the frame level quality scores but this metric is less competitive for blurred and noisy videos [3].

Visual Information Fidelity (VIF) is based on Human Visual System (HVS) model and visual statistics.VIF is derived as mutual information between two quantities, such as the mutual information between the input and output of the HVS channel with no distortion and the mutual information between the input and output of the HVS channel with distortion. VIF model, natural images in wavelet domain as Gaussian Scale Mixtures. This Gaussian Scale mixture model can measure the statistical features of natural images. HVS is also modeled in the wavelet domain. This metric in the image can be extended to video [5].

Singular Value Decomposition (SVD) is used to measure loss of textures and structure from the image. This is done by measuring the distortion as singular values. This measurement is done block by block and the final

value is taken by average. For video sequences these are done on luminance and chrominance components with corresponding weights and the overall quality is calculated by taking the average across all frames [6].

Most of the perceptual Video Quality Measurement Metrics were developed based on the traditional image quality assessment methods. Traditional image quality assessment methods include 1) A pre-processing process 2) Channel decomposition 3) Error normalization, or masking and 4) Error pooling. In preprocessing process stage color space transformation, image alignment, filtering and point wise transformations are done. The decomposition generally is done by Discrete Cosine Transform (DCT), Fast Fourier Transform, and Wavelet etc. This transforms the image signals into different spatial frequency as well as orientation selective subbands. Perceptual filters are added to extract the Spatio-Temporal information from the decomposed data. Spatio - Temporal sub regions are selected and from these quality features are extracted. These features are functions of space and time. By comparing features extracted from the calibrated processed video with features extracted from the original video, a set of quality parameters can be computed that are indicative of perceptual changes in spatial, temporal, and chrominance properties of video streams and these features are extracted from spatio-temporal (S-T) sub-regions using a mathematical function(e.g., standard deviation). Finally, a perceptibility threshold is applied to the extracted features. All features operate on frames within a calibrated video sequence. These perceptual based metrics are classified into frequency domain and pixel domain. In the frequency domain, transforms such as DCT, wavelets, and Gabor filter banks are used to measure the impairments in different frequency regions. In pixel domain, impairments are measured using change in local gradient strength around a pixel or based on perceptually significant visual features [7],[8].

This paper proposes an algorithm using wavelet and statistical measures that can exploit the perceptual property of Human Visual System with higher accuracy, low complexity than MS-SSIM and V-SSIM.

III. ALGORITHM FRAMEWORK

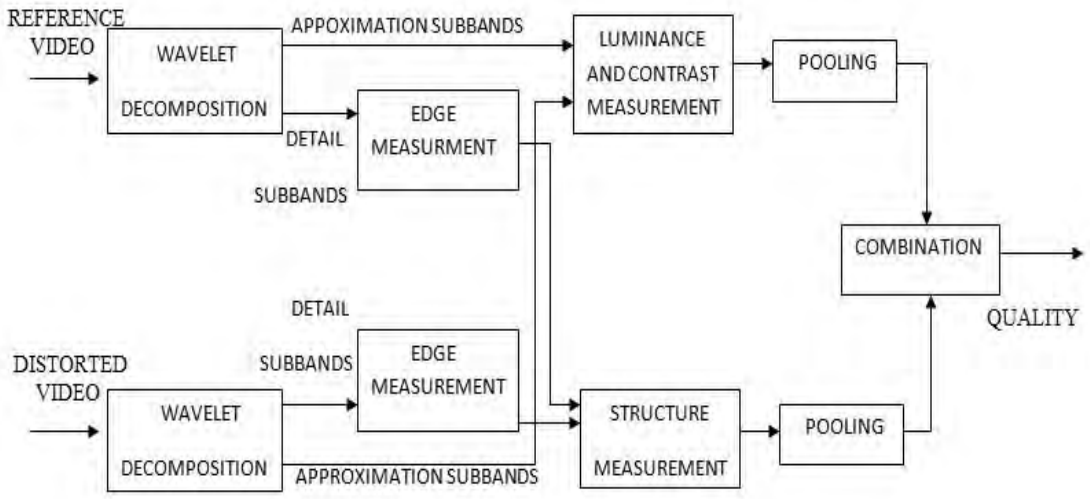

Figure:1 Block diagram of algorithm framework

The pre-processing processes such as color space conversion, alignment of the frames, filtering are done to reference and distorted videos. Color space used commonly for video processing is YUV. YCbCr is a color space similar to YUV. In YCbCr original video is in 4:4:4 sampling format. This original video signal requires large space for storage and it requires large bandwidth for transmission. So compression is done to the original video, based on exploiting the perceptual property of Human Visual System. But this introduces distortion in the video. The compression is done by Sub sampling the chroma signals and yields formats such as 4:2:2, 4:2:0. According to the amount of compression, distortion level also increases.

The output of the pre-processing process undergoes channel decomposition. In this paper channel decomposition is done by Haar wavelet, exploiting the features of Haar wavelet such as 1) shorter filter length, Its short filter length reduces computation and minimizes edge effects at image borders. 2) Integer implementation of this filter can be trivially obtained by setting the filter taps to have magnitude 1. 3) Any continuous real function on [0, 1] can be approximated uniformly on [0, 1] by linear combinations of the constant function 1, and their shifted functions. Wavelet transform exploits both the spatial and frequency correlation of data by dilations (or contractions) and translations of mother wavelet on the input data and perform spatial frequency decompositions. It supports the multiresolutional analysis of data i.e. it can be applied to different scales according to the details required. Characteristics of wavelet are well suited for compression

including the ability to take into account of Human Visual Systems (HVS) characteristics. Wavelet transform divides the information of an image into approximation and detail subbands. The approximation sub signal shows the approximated general pixel values and detail sub signals such as horizontal , vertical, and diagonal subsignals gives horizontal, vertical and diagonal detail coefficients [9],[10].

Filters are used to highlight or suppress features in an image subsignals based on spatio-temporal frequency. In this paper temporal variation is not considered. Spatial filters are used for suppressing noise or highlighting specific image characteristics. Here Gaussian filtering is used, and it is more suitable for removing Gaussian noise. Filtered components from the spatial filters are used for quality evaluation by using mathematical function such as mean, standard deviation. These quality features are functions of space and time. By comparing features extracted from the calibrated processed video with features extracted from the original video, a set of quality parameters can be computed that are indicative of perceptual changes in video quality. All features operate on frames within a calibrated video sequence.

Let X and Y are the pixel values from the reference and distorted video frames. Let $X_{A,}$ X_H, X_V and X_D are the approximation, horizontal, vertical and diagonal detail coefficients of the reference video frames. Similarly $Y_{A,}$ Y_H, Y_V and Y_D are the approximation, horizontal, vertical and diagonal detail coefficients of the distorted video frames. Luminance and contrast measurements are done on the approximation coefficients. Luminance measurements is calculated by using the equation

$$L(X_A, Y_A) = \frac{(2\,\mu_X\,\mu_Y + C_1)}{(\mu_X^2 + \mu_Y^2 + C_1)} \quad (1)$$

Where μ_x and μ_y denote the mean luminance intensities of the signals X and Y. For an image with a dynamic range *L*, the stabilizing constant is set to $C_1 = (K_1\,L)^2$ where K_1 is a small constant such that C_1 takes effect only when $(\mu_x^2 + \mu_y^2)$ is small.

Contrast measurements is done by

$$C(X_A, Y_A) = \frac{(2\,\sigma_X\,\sigma_Y + C_2)}{(\sigma_X^2 + \sigma_Y^2 + C_2)} \quad (2)$$

With σ_x and σ_y denoting the standard deviations of the luminance samples of the two images and C_2 is a stabilizing constant similar to C_1.

Detail subbands such as horizontal, vertical and diagonal elements are used for finding the edge measurements. Let μ, λ, and ψ be the coefficients of horizontal, vertical and diagonal detail subbands. Edge of reference frames denoted by $X_{E,}$ can be calculated by using the equation

$$X_E = \sqrt{(\mu(X_H^2) + \lambda(X_V^2) + \psi(X_D^2))} \quad (3)$$

Similarly Edge of distorted video frames denoted by $Y_{E,}$ can be calculated by using the equation

$$Y_E = \sqrt{(\mu(Y_H^2) + \lambda(Y_V^2) + \psi(Y_D^2))} \quad (4)$$

Human Visual System (HVS) is highly sensitive for horizontal and vertical components and less sensitive for diagonal components. Here arbitrarily choose $\mu = \lambda = 0.45$ and $\psi = 0.10$ under the condition $\mu + \lambda + \psi = 1$. Furthermore, structure comparison function is done with the covariance of the edge luminance samples σ_{xy} as

$$S(X_E, Y_E) = \frac{(\sigma_{XY} + C_3)}{(\sigma_X\,\sigma_Y + C_3)} \quad (5)$$

Contrast, Structure and Luminance measurements, calculated by using the above equations are averaged to get the quality of video frames called spatial pooling process. After the spatial pooling process these measurements are multiplied to get quality of a video frame denoted as Q(X,Y)

$$Q(X,Y) = L(X_A, Y_A) \,.\, C(X_A, Y_A) \,.\, S(X_E, Y_E) \quad (6)$$

This method is applying in frame-by-frame on the luminance component of the video and the overall quality of the video is computed as the average of the frame level quality scores. This process is called temporal pooling.

This newly proposed algorithm satisfies the following conditions:

1. Symmetry : $Q(X,Y) = Q(Y,X)$;
2. Boundedness : $Q(X,Y) \leq 1$;
3. Unique maximum : $Q(X,Y) = 1$ if and only if $X = Y$.

IV. IMPLEMENTATION AND RESULT

To implement the above proposed method, read original and distorted videos which are in avi format. Decompose the videos into frames, then convert into YCbCr format and perform spatio-temporal alignment. After the Spatio–Temporal alignment channel decomposition are done by using Haar wavelet. Extracting the features by point wise operation is not possible because it does not consider the neighbouring pixels. Most features such as edge, texture influence the neighbouring pixels. So to enhance the features and to remove noise a spatial filter such as Gaussian Low Pass Filter is used. This removes the Gaussian noise and enhances the features by standard convolution operation. Luminance, contrast and structure are calculated using statistical measures such as mean, variance and covariance on the filtered subband signals. These statistical measures are combined together to get quality of a frame. Quality from frames are pooled together to get overall quality. Performance comparisons are done with the existing algorithms such as MSE/PSNR, SSIM, MS-SSIM and V-SSIM. The proposed algorithm performs better than PSNR, SSIM, MS-SSIM, V-SSIM and MS-SSIM.

TABLE: 1 SHOWS QUALITY DERIVED BY THE CORRESPONDING METRICS.

METRIC	QUALITY	TIME DURATION (sec)
MSE	13.1672	0.0907
PSNR	37.0892	0.1099
SSIM	0.9746	5.7927
MS-SSIM	0.9780	4.0627
V-SSIM	0.9792	17.2783
PROPOSED ALGORITHM	0.9816	5.4374

V. CONCLUSIONS

As current video technology is increasing day to day life, assessing the quality of video is very important in wireless video applications. Traditional point based metrics are low complexity method but not correlating well with the human perception. In this newly proposed algorithm decomposition is done by Haar wavelet, which can gives finest features of the video frames. Since the Haar wavelet has the property of arranging it as a linear combination of constant functions and their shifted versions, the quality evaluation can be done by statistical analysis method. This newly proposed algorithm is compared with the MSE/PSNR, SSIM, MS-SSIM and V-SSIM. Compared to them proposed algorithm gives high accuracy, low complexity and it can track more the perceptual distortions.

REFERENCES

[1] Kalpana Seshadrinathan, Rajiv Soundararajan, A. C. Bovik, and Lawrence K. Cormack , “Study of Subjective and Objective Quality Assessment of Video“, IEEE transactions on image processing, vol.no.2009.

[2] Shyamprasad Chikkerur, Vijay Sundaram, Martin Reisslein, and Lina J. Karam, “Objective Video Quality Assessment Methods: A Classification, Review, and Performance Comparison”, .IEEE transactions on broadcasting, vol. 57, no. 2, june 2011.

[3] Zhou Wang, Ligang Lu and Alan C. Bovik, “Video Quality Assessment Based on Structural Distortion Measurement “, signal processing: image communication, vol. 19, no. 1, january 2004.

[4] Zhou Wang, Eero P. Simoncelli1 and Alan C. Bovik, “multi-scale structural similarity for image quality assessment” , IEEE asimolar conference on signals,systems and computers, 2003.

[5] Hamid Rahim Sheikh, Alan Conrad Bovik, and Gustavo de Veciana, “An Information Fidelity Criterion for Image Quality Assessment Using Natural Scene Statistics”, IEEE transactions on image processing, vol. 14, no. 12, december 2005.

[6] Manish Narwaria and Weisi Lin, “objective image quality assessment with singular value decomposition”, IEEE 2005.

[7] Stephen Wolf ,Margaret Pinson,” Video Quality Measurement Techniques” , NTIA Report,02-392.

[8] Arthur A. Webster, Coleen T. Jones, Margaret H. Pinson, Stephen D. Voran, Stephen Wolf, “An objective video quality assessment system based on human perception”, institute for telecommunication sciences,national telecommunications and information administration,325 broadway, boulder, co 80303.

[9] Kamrul Hasan Talukder and Koichi Harada, “Haar Wavelet Based Approach for Image Compression and Quality Assessment of Compressed Image” , IAENG international journal of applied mathematics, feb- 2007.

[10] S. Rezazadeh, S. Coulombe, “Novel discrete wavelet transform framework for full reference image quality