Choosing and implementing an open and free to use Objective video quality assessment tool

Kristian Skarseth
Department of Informatics,
University of Oslo,
kriskars@ifi.uio.no

September 3, 2014

Abstract

1 Introduction

Today there are no open source and free to use implementations of an accepted standard for full reference objective video quality assessment. In this paper we present the preliminary work for what will be an open source and free to use implementation of OPTICOMs PEVQ model [1]. We will also provide an introduction to the history of digital video quality assessment, including validation test results for various full reference (FR), reduced reference (RR) and no reference (NR) assessment methods which is the basis for our reasoning behind choosing the PEVQ model. Finally we present the current progress of our implementation of PEVQ, including some preliminary results and an outline for future work.

1.1 Background

From the time digital video codecs became widely used in the early 1990s there has been a demand for ways to objectively assess the quality of encoded video. Relying on subjective quality assessment is both time consuming and expensive, while a tool for objective quality assessment would be able to reduce both the cost and the time of measuring the quality of the video. The necessity for such a tool has been acknowledged by many, and today several standards for objective video quality assessment have been written.

1.2 Problem Definition and goals

As we will show in this paper, much work has been done towards creating an objective video quality assessment tool that can be as accurate as a well designed subjective test. The latest of such tools that provide the best results and resolution support are implemented and copyrighted by various companies, making it too expensive for most researchers and other potential users to benefit from using the software. Instead these users are forced to use less accurate tools which may be cheaper or free. This has motivated us to find and create our own free to use and open implementation of the best available open standard. It is also our goal to submit our solution to VLC [2] and have it incorporated in their group of video handling software in order to make it easily and widely accessible.

1.3 Outline

In section 2 we present the history behind video quality assessment. Section 3 describes the various objective quality assessment approaches in detail. Section 4 provides the

arguments and reasoning behind choosing the PEVQ model, as well as a brief explanation of the model itself. Section 5 showcases the current implementation and preliminary results, and section 6 presents a summary of the paper, as well as a separate summary of possible future work.

2 History

Following digital video codecs becoming more and more mainstream in the early 1990s, a big challenge was, and still is, to create codecs that provide the optimal ratio of video quality and data quantity, and to find the best way of streaming digitally encoded video sequences over the Internet. For codecs, the correct ratio will depend on the use case. You may accept a higher data quantity for an increased video quality when you store the original of a video clip. However you may also have to live with lower quality in order to reduce the data quantity when streaming the video over the internet.

Testing the quality of the video, especially after the video sequence comes out on the other end of a streaming procedure, is an important step in both finding the correct codec and especially the correct streaming protocol. Producing a subjective test score relies on well designed and organized tests with several test subjects. Performing such a test will likely be both expensive and time consuming. On the other hand, a tool for producing an objective quality assessment score could be able to rate the quality of the video faster, cheaper and without the involvement of large subjective testing groups. Researchers with access to such a tool would be in a good position when testing anything they are developing that handles and/or alters video sequences.

2.1 Peak Signal to Noise Ratio

A much used approach to testing video quality objectively is Peak Signal to Noise Ratio (PSNR). It is simple, and has been used for a long time. However, when the original video content is altered in any way, the quality measured by PSNR is not reliably close to the assessment results coming from subjective testing. A better approach is taking the human visual system (HVS) into account when analysing the video quality. This is explained, among others, by Huynh-Thu and M. Gihanbari in [3] and Bernd Giron in [4].

2.2 Validation of Perceptual Objective Video Quality Assessment standards

Unlike PSNR, perceptual objective video quality assessment tools attempt to mimic the HVS so that the results may be closer to subjective test results. In order to meet require-

ments and expectations from users for a potential objective video quality assessment tool, a subcommittee of The American National Standards Association (ANSI), named Accredited Alliance for Telecommunications Industry Solution (ATIS), performed a validation test for objective video quality measurement. As described by Pinson, Staelens and Webster in [5], the report [6] from the testing, performed by the group T1A1.5 within ATIS, laid the groundwork for future tests of objective video quality assessment tools. Two standards were also formed after the testing, the ANSI Standard T1.801.03 and T1.801.01. The following information is based on a paper by Pinson, Staelens and Webster [5], summarizing the history of video quality model validation.

In 1997 the first meeting of the Video Quality Experts Group (VQEG, www.vqeg.org) found place in Turin. VQEG was formed so that international experts within subjective video quality measurement could come together and share their information. The purpose of the group was and still is to advance within the field of video quality assessment.

To date the VQEG group has gone through several large testing phases. From 1999 to 2000 their first phase, called the full reference television (FRTV) phase I, were conducted by the Independent Lab Group (ILG). It was designed for testing objective full- and no-reference standard definition television quality; however none of the no reference models made it to the testing phase. The conclusion from the test was that none of the submitted models were statistically better than PSNR.

Following the FRTV phase I came FRTV phase II (2002-2003), the multimedia phase I (2007-2008), reduced reference/no reference television (RRNR-TV) phase I (2008-2009) and the high definition television (HDTV) test (2009-2010, all tests conducted by ILG with some proponents involved in certain cases. Eight full reference FR models were published in a first rendition of ITU-T Rec. J144 following FRTV phase I (2001), and FRTV phase II published a revised version of ITU-T Rec. J144 as well as ITU-T Rec. BT.1683, where two FR models were standardized. In both phases all no reference (NR) models were withdrawn. Following the multimedia phase I, FR models from Nippon Telegraph and Telephone Corporation (NTT), OPTICOM, Psytechnics and Yonsei University were standardized in ITU-T Rec. J.247 and ITU-R BT.1866. A reduced reference (RR) model from Yonsei University was standardized in ITU-T Rec. J.246 and ITU-R BT.1867. Again no NR models were standardized. The RRNR-TV phase I test standardized 3 RR models in ITU-T Rec. J.249, and the HDTV test in 2009-2010 standardized two FR models in J.341 and a RR models in J.242. Two NR models were mentioned in VQEGs final report for the HDTV test, but neither was standardized. For more information on the history of VQEGs validation tests, see Staelens and Websters paper in [5].

Table 1 shows a summary of the test phases we have talked about with the name of the test phase, name of the organization who designed the test, the date, tested resolutions and standards documents for objective video quality assessment published following the tests.

Table 1: Overview of test phases

Test-phase name	Org.	Date	Resolutions	Standards documents
T1A1	ATIS	1994-1995	NTSC	T1.801.03 & T1.801.01
FRTV Phase I	VQEG	1999-2000	NTSC & PAL	ITU-T Rec. J.144
FRTV Phase II	VQEG	2002-2003	NTSC & PAL	ITU-T Rec. J.144 & ITU-R Rec. BT.1683
Multimedia	VQEG	2007-2008	VGA, CIF & QCIF	ITU-T Rec. J.247, ITU-R BT.1866, ITU-T Rec. J.246 & ITU-R BT.1867
RRNR-TV Phase I	VQEG	2008-2009	NTSC & PAL	ITU-T Rec. J.249
HDTV	VQEG	2009-2010	1080i & 1080p	ITU-T Rec. J.341 & ITU-T Rec. J.242

In order to further improve and advance the field of validation of video quality metrics, VQEG have also started the Joint Effort Group (JEG) which is developing tools and laying a groundwork for others who wish to validate video quality metrics. More about the JEG group can be read in Staelens et. al. paper [7].

In the next section we will provide an introduction to exactly what objective video quality assessment means, and the difference between the three approaches mentioned in this section, no-, reduced- and full-reference objective video quality assessment.

3 Objective video quality assessment

As we have been able to see from the history of objective video quality assessment, the approach with highest consistency compared to the subjective counterpart is the full reference (FR) version. This is however also the approach that requires the most background data for any video sequence in order to give results. The reduced reference (RR) and no reference (NR) methods can be executed without access to the full source material. In order to better understand our decision to choose a FR model the next three sections will give an overview of the three approaches to objective video quality assessment. To read more about the differences to the methods mentioned in the following three sections you can see S. Chikkerur and V. Sundarams paper in [8].

3.1 No reference objective video quality assessment

NR points to the fact that there is no reference to the original source material when measuring and determining the quality of the video. This means that the assessment

is done purely by analysing the degraded signal. In an environment where there is no access to any source material the only possible objective approach will be a NR method. However as we could see in section 2 the performance of NR methods are to this date significantly lower compared to that of RR and FR approaches.

3.2 Reduced Reference objective video quality assessment

RR is a middle ground between NR and FR in that it has access to certain information about the source signal. The information can be transmitted where there is limited transmission capacity and therefore limited access to the reference signal. This limited extra information about the source signal can still be enough to provide better results than a NR approach, and as we saw in section 2 several standards have been approved for RR methods.

3.3 Full Reference objective video quality assessment

FR is the final approach where the entire source signal is available together with the degraded signal. This means we can compare the two signals as detailed as on a per pixel basis and draw a conclusion from the difference between the two signals. Considering all data is available during the measurement it makes sense that a FR technique will be most accurate, and as we could see from section 2 this has also been the case in tests designed by VQEG.

4 Choosing the standard

It was important to us to choose an approved standard when implementing an open and free to use solution. If anyone is to rely on the tool we create they must be able to trust that it gives results at least close to that of other tools you can purchase. Selecting one of the models standardized through VQEG testing was therefore a natural choice. After taking that decision we had to choose between a NR, RR and FR model. To date no NR models have been standardized, as shown in section 2, which makes it easy to dismiss any such models. The choice between RR and FR models has come came down to choosing the open standard that can be implemented without paying royalties, while also having performed very well in testing. We have chosen OPTICOMs PEVQ model which was standardized in ITU-T Rec. J.247 [1]. Three other models were also available in this standard, but the OPTICOM model scored overall best which can be seen in table 2, 3 and 4. The tables use values from the J.247 standard, where the correlation values represent how closely the results from each model correlates with subjective test results. The closer to 1 the correlation value is, the better is the result.

Table 2: VGA resolution

VGA resolution	NTT	OPTICOM	Psytechnics	Yonsei	PSNR
Avg. correlation	0.786	0.825	0.822	0.805	0.713
Min. Correlation	0.598	0.685	0.565	0.612	0.499

Table 3: CIF resolution

CIF resolution	NTT	OPTICOM	Psytechnics	Yonsei	PSNR
Avg. correlation	0.777	0.808	0.836	0.785	0.662
Min. Correlation	0.675	0.695	0.769	0.712	0.440

The biggest limitation of the PEVQ model from ITU-T Rec. J.247 is the tested resolutions which are limited to VGA, CIF and QCIF. HD resolution is Today in full use in most applications, and choosing a standard which is not designed for HD seems strange. This is unfortunately a limitation we have to accept as the models described in newer standards, such as ITU-T Rec. J.341, that does support HD, have been patented in such a way that we cannot make our own implementation for free.

4.1 Description of PEVQ

PEVQ as standardized in ITU-T Rec. J.247 is a FR objective video quality assessment model designed to measure video quality on mobile and multimedia platforms. It does so by calculating five indicators, each operating in different domains (temporal, spatial, luminance and chrominance) motivated by the human visual system (HVS) so that it may attempt to see the video as a human subject would. The model requires two input signals, the full reference (source) signal and the full degraded signal. The model will go through several stages where both signals are processed and results from the processing stages are weighted together to provide the final result.

Using figure 1 we describe the basic flow of PEVQ. The first stage consists preprocessing where a spatial region of interest (RIO) is extracted from the two signals. The RIO is the area on the video where all the following processing will be performed, while the leftover edges are used only for certain calculations where neighbour pixeldata is required. Following the RIO extraction comes a coarse alignment registration of the input sequences which in step 3 is used to perform the detailed sequence alignment that compensates for shifts in the spatial domain. The resulting values from step 3, named “matching info”, are used to find the perceptual impact of temporal degradations. We will also have the final cropped and aligned versions of the two signals after this stage.

In stage 4 the perceptual difference in the spatial domain between the two signals is processed, providing four distortion indicators. The matching info from stage 3 is

Table 4: QCIF resolution

CIF resolution	NTT	OPTICOM	Psytechnics	Yonsei	PSNR
Avg. correlation	0.819	0.841	0.830	0.756	0.662
Min. Correlation	0.711	0.724	0.664	0.587	0.540

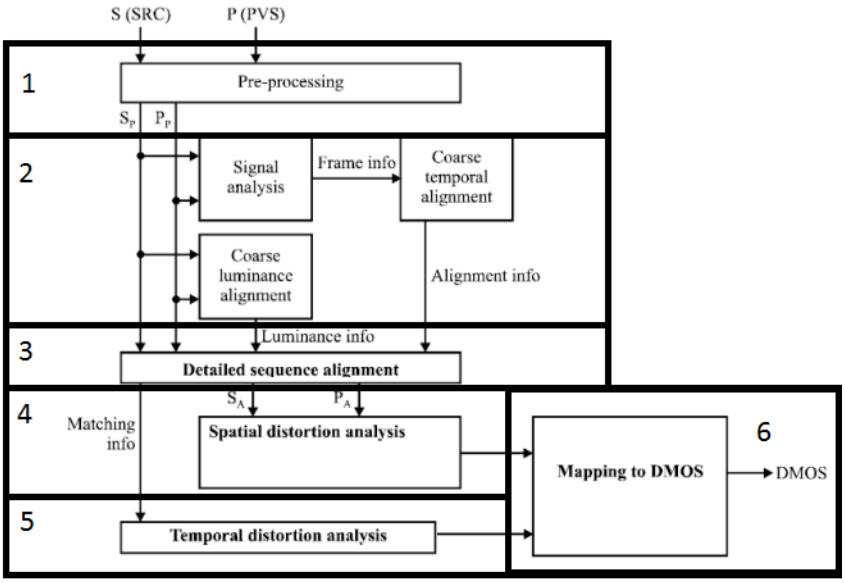

Figure 1: Overview of the PEVQ model.

further analysed in stage 5 to create one indicator that represents temporal distortions.

In stage 6 the five main indicators from the previous calculations are weighted together to create the final score. As described in the standard, the final result correlates highly with a mean opinion score (MOS) obtained from subjective tests. It is important to remember that MOS is a widely used term, and there is no definition on what range of values a MOS score must be within. MOS simply represents a scale with which we can give a meaningful score. A description of various ways to use MOS can be found in [9]. To get a deeper understanding of how PEVQ does its analysis, please see the J.247 publication in [1].

5 Implementation

We have now provided a brief history of objective video quality assessment, explained the various approaches one may take in assessing video quality and provided reasoning for choosing the standardized model we have. In this section we will explain the decisions made with regards to the implementation itself. We will also present the

current progress of our implementation, and talk about the work laying ahead of us.

5.1 Coding decisions

Since the PEVQ model from ITU-T Rec. J.247 is designed for resolutions up to VGA we hope to have the opportunity to further develop and test the implementation to support high definition resolutions. As we want to focus on the PEVQ implementation itself rather than the video handling surrounding it, it is important to choose the shortest path towards writing PEVQ implementation code. Being able to not create a whole framework for video decoding, encoding, writing etc will save us a lot of valuable time. This may also provide us with the time to further build on the PEVQ model to support resolutions higher than VGA in the future. in order to skip this step we have decided to use the open source library, Libav [10] which allows us to start implementing algorithms from PEVQ immediately after setting up an environment for Libav.

Libav is a C library with bindings to several other languages, but as the algorithms in PEVQ require a high amount of computing power it is essential to choose a coding language where high performance is possible. Writing the implementation in C or C++ was therefore a natural choice, whereas we have decided to focus on C++ to create a modern and object oriented solution.

We would also like to briefly mention the possibility of porting the program to a graphics processing unit (GPU). While this is not part of the initial plan, we are aware of the speedup possibilities that lie within GPU execution when dealing with image or video processing the way we are. One of the more computing intensive parts of the PEVQ solution is the calculation of the edge images using a Sobel Filter. A demonstration of speedup gained when executing a Sobel edge filter on a GPU rather than a CPU can be viewed in a paper by A. Dore and S. Lasrado in [11]. Allowing users to run the video quality assessment tool faster will surely be an appreciated feature.

As may be interpreted from section 4.1, the PEVQ model itself is somewhat modular. Each step is more or less standalone, relying on the results from a previous step in varying degrees. Before the actual quality assessment can be performed, the model makes sure the source and degraded signal are aligned properly. This is so that when comparing results the comparison is done between the correct parts of the video in case the degraded signal has been stretched in any way. Should the analysis stage learn that the video sequences are satisfactory aligned, the alignment module of the code can be skipped. We will design our solution in such a way that if the assessment tool detects something which can be skipped, it can proceed with the next step and get a performance gain.

5.2 Preliminary results

So far we have been able to implement code for the pre-processing and signal analysis stages, in addition to Libav specific code to handle loading, encoding, decoding and writing of video files to disk. This means we are able to load any video file, decode it into a YUV format we can work with in our PEVQ implementation, for later to write video sequences for testing purposes to disk. We have also created simple data dumping for values created in the signal analysis stage which allows us to control the results.

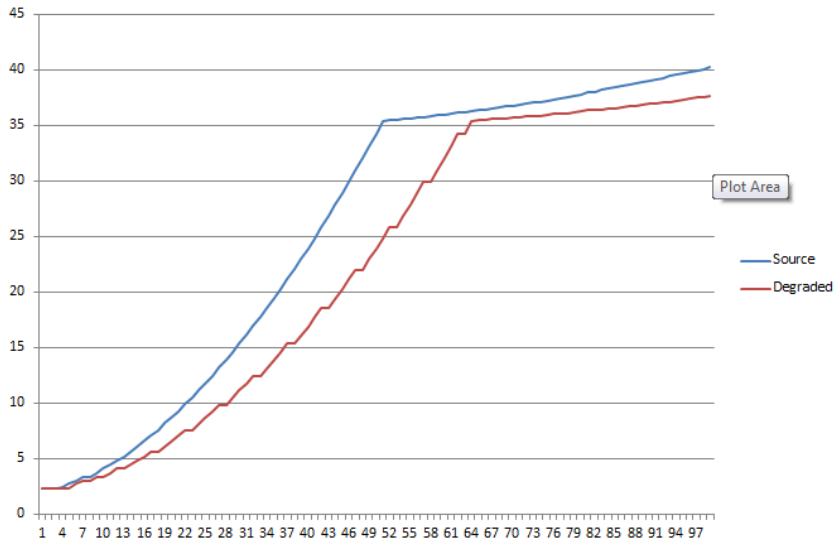

Figure 2: Source and degraded standard deviation edge image.

Figure 2 visualizes the values from the standard deviation of the Sobel Filter edge image. In this test we used a source video sequence at 24 frames per second and a degraded signal at 30 frames per second. We see that the degraded signal is lagging behind the source signal more and more, and that the standard deviation value is completely similar for two consecutive frames each time the degraded signal seems to be additionally delayed. While this is just a small part of the PEVQ model, the standard deviation from the edge image, among other results, is used to perform the coarse temporal alignment as described in section 4.1. The result from figure 2 clearly illustrates that the video sequence will need some sort of temporal alignment in order for the two video sequences to be properly aligned. If such an alignment is not performed, further analysis on a frame by frame basis will not match in the time domain. With regards to what we talked about in section 5.1, this discovery translates to the fact that we need to run the alignment module.

6 Summary

We have described today's situation with regards to objective video quality assessment, where available tools are either too expensive for most users, or provide results that correlates poorly with subjective test results. We have given an overview of the history of standardization of no-, reduced- and full-reference objective video quality assessment tools as well as a description of the difference between these approaches. Based on the test results of the standardized models we have argued for why we have chosen the PEVQ model which we will implement, and we have given a brief description of design decisions and plans for our own implementation. Finally we have given an overview of our current progress on the implementation, including a description of the results it provides when testing a video sequence with certain degradation.

6.1 Future work

The main goal and what we will begin working on is implementing the PEVQ model exactly as it is described in the standardization ITU-T Rec. J.247. This work has begun, and will continue over the last half of 2015. As mentioned in section 4 the standardized PEVQ model we are implementing does not support HD resolution. Another research topic we must keep in mind once our implementation is done is therefore extending our solution to support resolutions above VGA. We also mentioned in 5.1 that speedup using a Graphics Processing Unit(GPU) can potentially be significant, and we would like to look at the possibility of performing parts of the calculation in the PEVQ model on a GPU.

References

[1] International Telecommunication Union. Objective perceptual multimedia video quality measurment in presence of a full reference, 2008.

[2] VideoLan. Videolan organization. `http://www.videolan.org/`. [Online; last checked 31.05.14].

[3] Q. Huynh-Thu and M. Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics Letters*, 44(13):800–801, June 2008.

[4] Bernd Girod. Digital images and human vision. chapter What's Wrong with Mean-squared Error?, pages 207–220. MIT Press, Cambridge, MA, USA, 1993.

[5] M.H. Pinson, N. Staelens, and A. Webster. The history of video quality model validation. In *Multimedia Signal Processing (MMSP), 2013 IEEE 15th International Workshop on*, pages 458–463, Sept 2013.

[6] Coleen Jones, Ned Crow, Stephen Wolf, and Arthur Webster. Analysis of t1a1.5 subjective and objective test data, 1994.

[7] N. Staelens, I. Sedano, M. Barkowsky, L. Janowski, K. Brunnstrom, and P. Le Callet. Standardized toolchain and model development for video quality assessment # x2014; the mission of the joint effort group in vqeg. In *Quality of Multimedia Experience (QoMEX), 2011 Third International Workshop on*, pages 61–66, Sept 2011.

[8] S. Chikkerur, V. Sundaram, M. Reisslein, and L.J. Karam. Objective video quality assessment methods: A classification, review, and performance comparison. *Broadcasting, IEEE Transactions on*, 57(2):165–182, June 2011.

[9] International Telecommunication Union. Itu-t recommendation p.800: Metods for subjective determination of transmission quality, 2008.

[10] Libav. Libav - about libav. `https://libav.org/`. [Online; last checked 28.05.14].

[11] Aruna Dore and Sunitha Lasrado. Performance analysis of sobel edge filter on heterogeneous system using opencl. *International Journal of Research in Engineering and Technology (IJRET)*, 3:53–57, 2014.