Education and Culture DG

ERASMUS MUNDUS

Master Erasmus Mundus in
Color in Informatics and Media Technology (CIMET)

TOWARDS A PERCEPTUAL METRIC FOR VIDEO QUALITY ASSESSMENT

Master Thesis Report

Presented by

Seyed Ali Amirshahi

and defended at the

Gjovik University College, Norway

8th June 2010

Jury Committee:	Supervisor(s):
Professor Alain Tremeau	Dr. Mohamed-Chaker Larabi
Dr. Damien Muselet	Professor Jon Yngve Hardeberg

Abstract

In the last couple of years a huge amount of work has been shifted from works on Image Quality Assessment to Video Quality Assessment. Although some metrics have started to take the temporal aspects of videos into account but still most metrics are focusing on the spatial distortions in videos or in other words applying image quality metrics on individual frames. Also till now most metrics are focusing on the Quality of Service (QOS) rather than the Quality of Experience (QOE). With respect to the mentioned factors we believe that there is a need for a new metric which has a Spatial-Temporal approach and takes QOE into account and so the proposed metric is based on these two main approaches. Because of the spatial-temporal approach we had the metric was named as STAQ (Spatial-Temporal Assessment of Quality).

Our proposed method is based on the fact that the Human Visual System (HVS) is sensitive to sharp changes in videos. Keeping this in mind we could reach the conclusion that there will be matching regions in consecutive frames. We took advantage of this point and found these regions and used a Full Reference Image Quality Metric to evaluate the quality of these frames. We also used five different Motion Activity Density groups to evaluate the amount of motion in the video. Our final score was later pooled based on five different pooling functions each representing one of the motion activity groups. In other words we used QOE or information from subjective evaluation for playing a controlling factor role in our method.

When the proposed reduced reference metric is compared to ten different state of the art full reference metrics the results show a great improvement in the case of H.264 compressed videos compared to other state of the art metrics. We also reached good results in the case of MPEG-2 compressed videos and videos affected by IP distortion. With respect to the results achieved we could claim that the metric introduced is among the best metrics so far and has especially made a huge progress in the case of H.264 compressed videos.

Preface

The research work presented in this thesis was carried out in the School of Computer Science at the University of Poitiers under the scholarship awarded to the author by the CIMET consortium. I am therefore, grateful to European Commission for giving me the opportunity to conduct the research work presented in this thesis.

First and foremost I would like to thank my supervisor Dr. Mohamed-Chaker Larabi for all the time he spent with me on my work and the great ideas he gave for improving the work. Obviously, finishing the work as it is would not have been possible without his guides and supports. I should also thank him not only for the time he put on my work but the time he spent during these couple of months listening to me and acting like an older brother which I didn't have around.

I should also thank my colleges here at the University of Poitiers, Dr. Aldo Maalof, Rafik Bensalma, Michael Nauge and Miryem Hrarti. Not only the comments they gave increased the quality of the work but they made the working environment more enjoyable.

Finishing this CIMET master program and the master thesis was impossible without the presence of the CIMET administrative coordinator, Mrs. Helene Goodsir. From a couple of months before the starting of the CIMET program which we were all applying for our visas till now that is the last days we have made her different types of troubles and so we own a lot for all the things she has done.

I also like to thank all my friends and teachers in the CIMET program which not only made this program a high standard program but also gave us the chance to get to know different cultures and countries.

Last but not least, I would like to emphasize my deepest gratitude to my family especially my parent which I own everything I have in my life to them. I hope I would be able to make it up to them in some way. Not only they supported me in any way they could during my master thesis but because of their academic background I got some really good advices from them as well.

Table of Contents

Table of Figures

1 Introduction

1.1 Introduction

Image and video quality assessment has long been an interesting field of research. Although a huge amount of work on Image Quality Metrics (IQM) have been done in the past decades [**1**], but Video Quality Assessment (VQA) is rather a new filed and there has been an increase on works focusing on it during the last couple of years [**2**]. It should be mentioned that great progress has been made in the field of Image Quality Assessment (IQA) and we can claim that it has more or less reached the point of maturity especially in the case of grey scale images. On the other hand VQA needs a lot of work to reach the point where IQM's are.

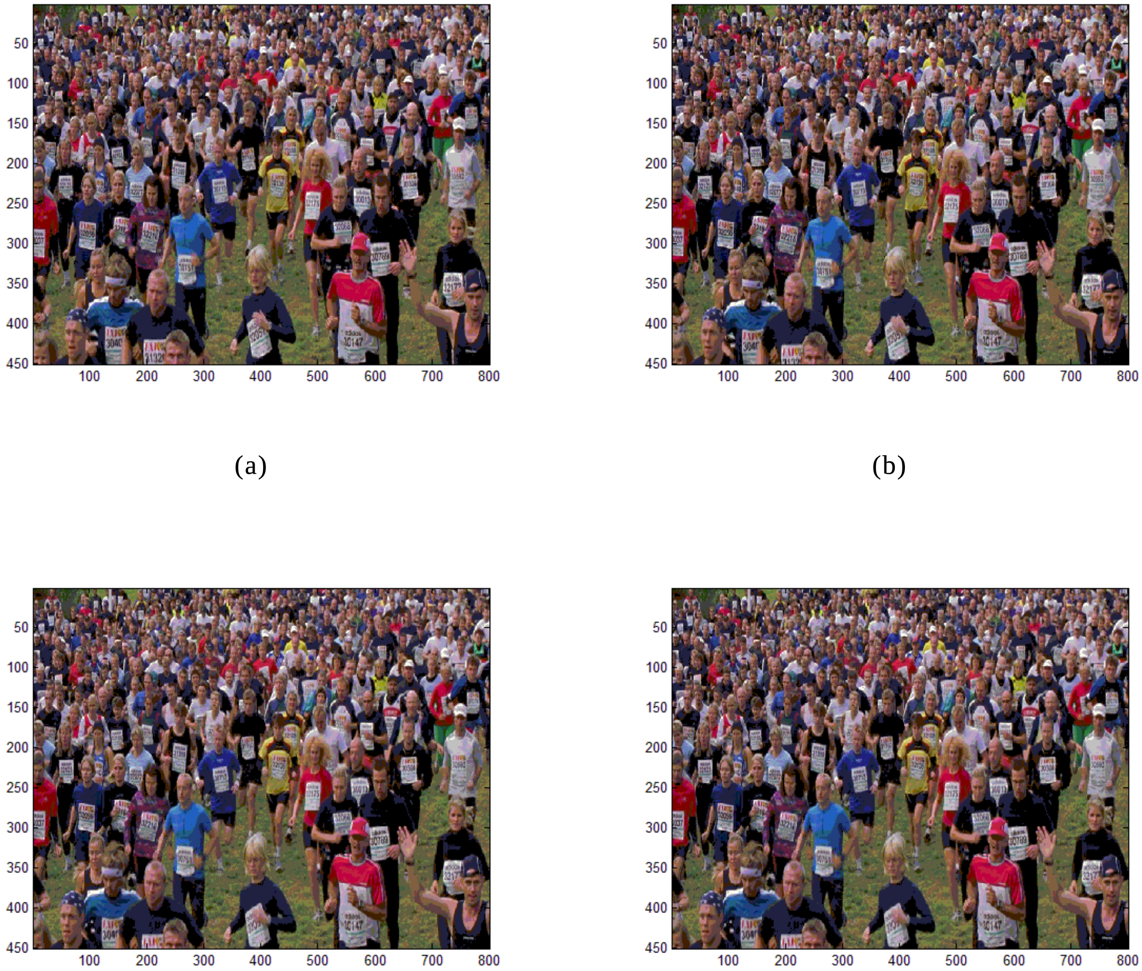
Videos can be seen everywhere from TV and cinemas which we were involved with for a long time to now days that video is more and more getting involved in our daily lives. Video on demand, video conferencing, Digital television and the videos we have access to on the internet are just a few examples of the videos we are dealing with, every day. It could be claimed that everyone is somehow in contact with videos in his/her daily life. This huge amount of video brings up the field of video processing. Video compression, video watermarking, and … are examples of some works done on videos. After any video processing work we should evaluate the quality of the output video to check if the output video has not lost its quality.

Unlike Image Quality Assessment Metrics (IQAM's) that we are only working in the spatial domain, in the VQM's we are working with both the spatial and temporal domain. VQM's which deal with the spatial domain mainly use an IQM to evaluate the quality of each frame and then pool the quality scores based on the different approaches. Although this might be an option but not taking the temporal domain into account is not a good idea. Some of the distortions we might face in the temporal domain are [**3**]:

1. Flicker.
2. Motion Inconsistency.
3. Mosquito Noise.
4. Spatio-Temporal Noise.

Another important factor on researches done in Video Quality Assessment Metrics (VQAM's) is a shift in trying to find metrics which are based on Quality of Service (QOS) to metrics which are based on Quality of Experience (QOE) [**4**]. QOS metrics have a systematical approach and use mathematical and signal tools to evaluate the quality of a video but metrics with a QOE approach try to bring the observer into the loop when they are calculating the quality. The reason behind this shift is because metrics with a QOE approach give results that have a higher correlation with results given by the subjective tests. This fact will for example help TV providers in giving a better service to their customers. Depending on the content of the video and how the viewers will react and evaluate the quality of it they can tune their system so that they would provide a better service to their customers.

The most important factor that an observer will take into account and is sensitive to when evaluating the quality score of a video is the amount of motion activity density along with the content of the video. An example of such an issue is that the Human Visual System (HVS) reacts differently to two different videos in two different motion activity groups but with the same type and level of distortion. A good example of two videos with different amount of distortion would be a news program, which would therefore have a small amount of motion activity by the news anchor and nearly no motion activity in the background and a video of people running which will have a high amount of motion activity. To give an idea about the difference between consecutive frames in videos with different amount of motion activity there are examples in Figure 1 to Figure 4. In Figure 1 which is a video from the SVT High Definition Test Set [**5**] we could see 4 consecutive frames of a video with a high amount of motion activity. In Figure 2 the difference between the consecutive frames shown in Figure 1 are shown, the colorbar next to the figures shows how different the frames are. In Figure 3, 4 consecutive frames of a video with a low amount of motion activity are shown and finally in Figure 4 the difference between the frames shown in Figure 3 is presented. As it can be seen depending on the amount of motion in a video the difference between consecutive frames changes.



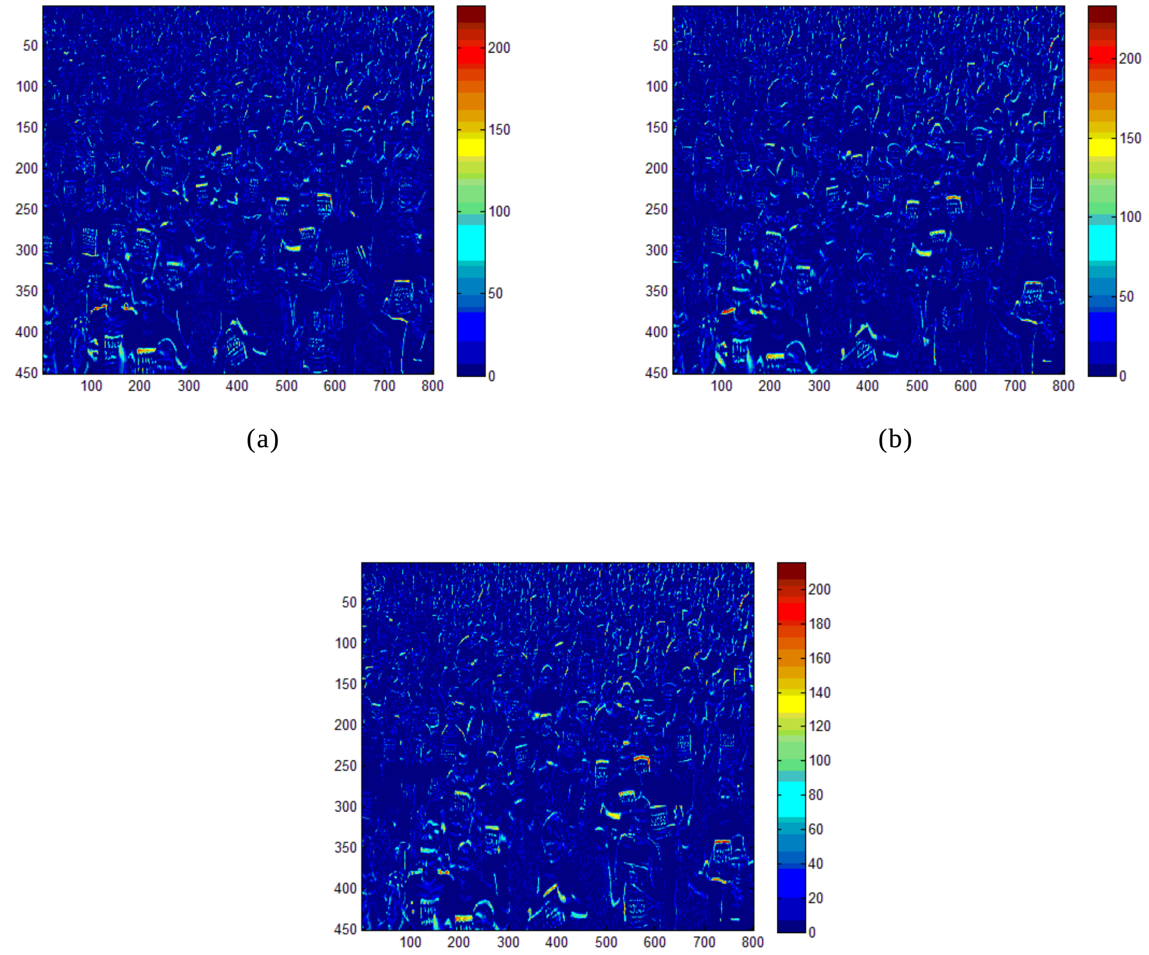
(a)

(b)

(c)

(d)

Figure 1. 4 consecutive frames of a video with a high amount of motion. Frames are ordered from (a) to (d).

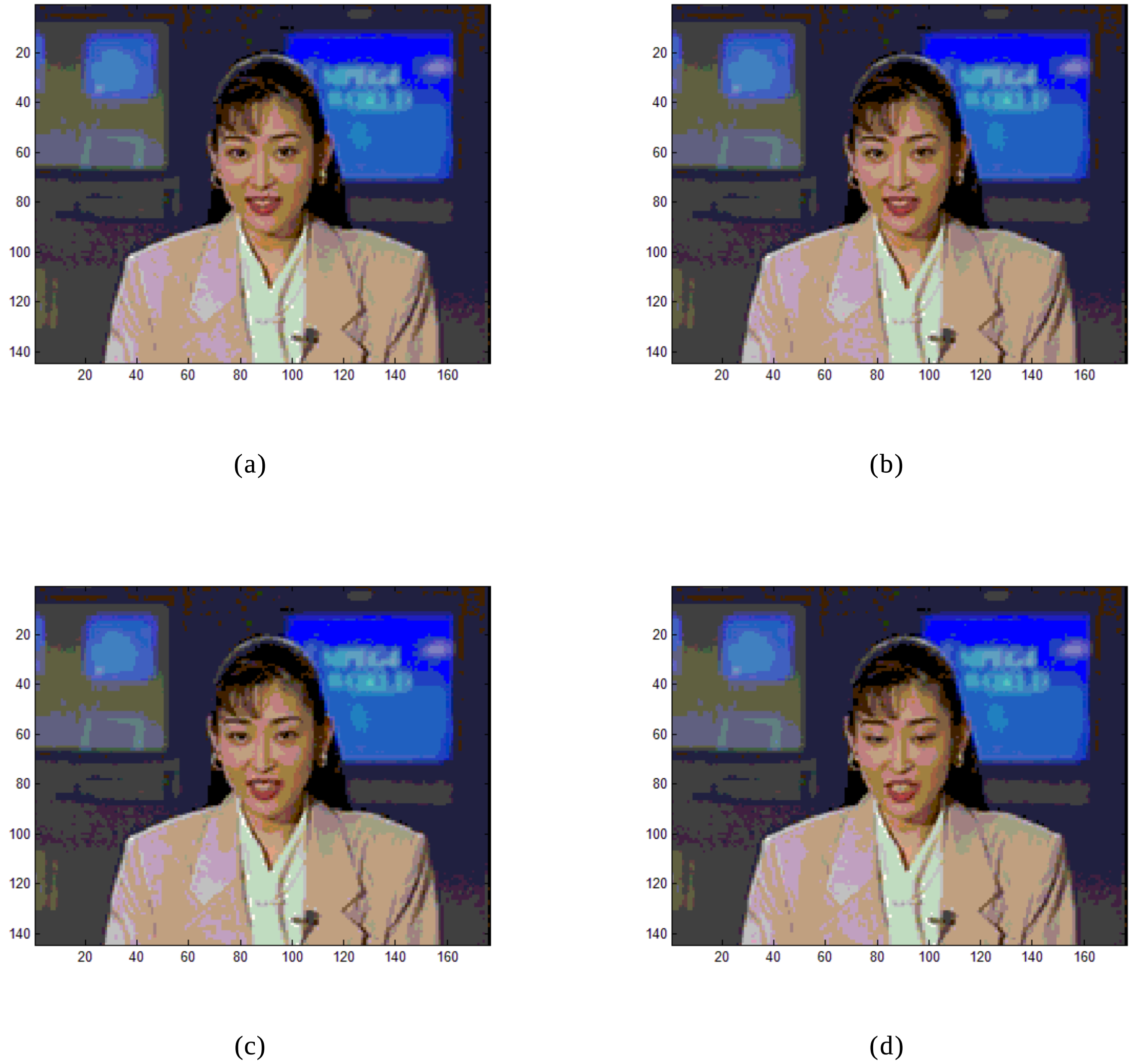


(a)

(b)

(c)

Figure 2. Difference between consecutive frames in Figure 1. (a) difference between Figure 1 (a) and 1 (b). (b) difference between Figure 1 (b) and 1 (c). (c) difference between Figure 1 (c) and 1 (d).



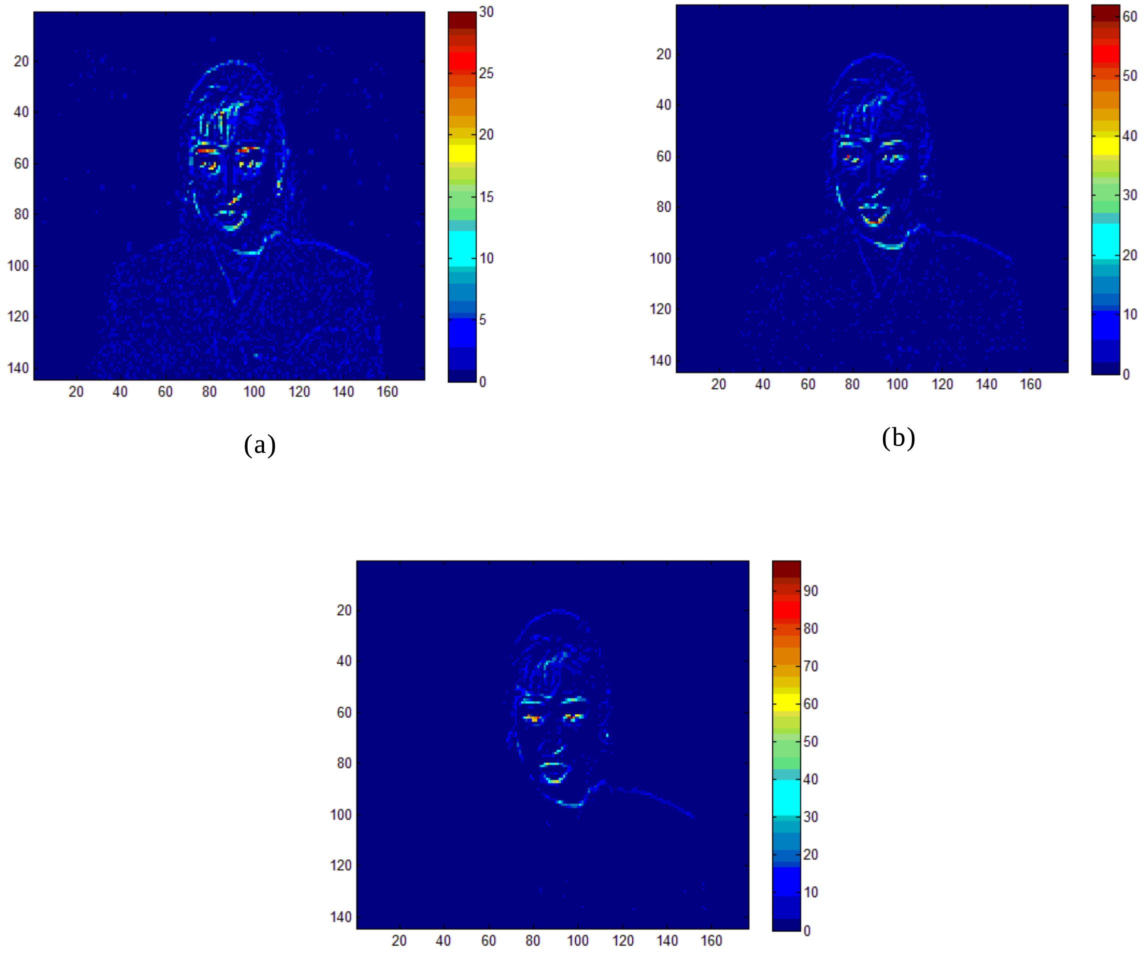
(a)

(b)

(c)

(d)

Figure 3. 4 consecutive frames of a video with a low amount of motion. Frames are ordered from (a) to (d).



(a)

(b)

(c)

Figure 4. Difference between consecutive frames in Figure 3. (a) difference between Figure 3 (a) and 3 (b). (b) difference between Figure 3 (b) and 3 (c). (c) difference between Figure 3 (c) and 3 (d).

1.2 Video quality assessment

Like IQA in VQA we have two main approaches Subjective assessment and Objective assessment of video quality.

1.2.1 Subjective quality assessment

In this approach we evaluate the video quality based on the observation and responses we get from human observers. Although using this method will give us accurate results which are based on the HVS but performing subjective tests have disadvantages as well. Like any other subjective test, performing subjective test in VQA is time consuming and financially expensive. Also there are a lot of pre-requirements to take into account before a subjective assessment is performed. These technical issues have been introduced in different standards such as [**6**]. For example the International

Telecommunication Union (ITU) has published a 48 page recommendation guide for subjective assessment of the quality of television pictures [**6**] which is also been adhered by the Video Quality Experts Group VQEG [**7**]. A number of the requirements needed for a subjective test in a laboratory environment are shown in Table 1.

Table 1. Some of the requirements needed to be taken into account for a subjective assessment in a laboratory environment [6].

Ratio of luminance of inactive screen to peak luminance.	≤ 0.02
Ratio of luminance of the screen, when displaying only black level in a completely dark room, to that corresponding to peak white.	≈ 0.01
Displaying brightness and contrast.	Set up via PLUGE
Maximum observation angle relative to the normal (this number applies to CRT displays, whereas the appropriate numbers for other displays are under study).	30
Ratio of luminance of background behind picture monitor to peak luminance of picture.	≈ 0.15
Chromaticity of background.	D65
Other room illumination.	Low

The requirements mentioned are just a small number of requirements which should be kept in mind when performing a subjective test. There are also a huge number of requirements regarding the observers, display and .... With regards to all these factors, we can see that although performing subjective tests for evaluating the quality of different videos is the ideal solution but it is hard or in other words impossible to keep up with all the requirements.

1.2.2 Objective quality assessment

Due to the reasons mentioned in section 1.3.1 running a subjective assessment test is not a convenient option. This is why objective assessment has been introduced. Objective VQM's try to model the HVS and give a single value (or in some rare cases a number of values) as the overall quality of the video. The correlation between the results coming from an objective metric and the subjective results show how good the metric is performing. Although some simple metrics such as Mean Square Error

(MSE) or Peak Signal to Noise Ratio (PSNR) have been introduced but due to the low correlation between their results and subjective results other complicated metrics have been introduced.

Depending on the amount of information we have from the reference video the metrics could be categorized in 3 different main groups:

1. Full Reference (FR) metrics.
2. Reduced Reference (RR) metrics.
3. No Reference (NR) metrics.

Among the current metrics introduced we can claim that the best results are reach when the FR metrics are used. This could be because although a huge amount of work and time has been put on studying the HVS but still there is a lot of work to do in this field.

1.2.2.1 Full reference metrics

In the FR methods both the reference video and the test video are available. In these methods the operation is made with respect to the reference video. Most metrics in VQA are categorized among this group [**8**] and this is a reason that the FR metrics are most widely used till today. Among different metrics proposed, MSE and PSNR are the oldest and most common metrics used. As it was mentioned in section 1.3.2 the results from these two metrics do not correlate well with the subjective test, the reason is that they do not take the HVS into account, further description on these two metrics will be presented in 2.2.1 and 2.2.2 With respect to all the negative points mentioned due to the simplicity of these two methods these methods are still used widely in VQA works.

1.2.2.2 Reduced reference metrics

In RR metrics we do not have the reference video itself but we still have some limited information/data regarding the reference video. RR metrics were introduced since transferring or having access to both the reference video and the test video is quite expensive and not always possible and in some cases we do not have access to the reference video itself. RR metrics are used in metrics that we have/need some information about the reference video but not all the reference video itself to evaluate the quality of the video.

1.2.2. 3. No reference metrics

NR metrics are metrics which evaluate the quality of a video without any prior knowledge about the reference video. It is a fact that human observers can rate the quality of a video without seeing its reference and just by observing the test video. But because of the limited information available regarding the HVS, there is not much work done on the NR metrics [**9**]. Mainly the NR metrics try to find different distortions in the frames and evaluate the video quality according to them. Blurring, blocking, quantization and ringing artifacts are some of the features extracted for this purpose [**10**], [**11**], [**12**] and [**13**].

1.3 Structure of the work

In this thesis we will review some of the proposed VQAM’s and the state of the art metrics in Chapter 2. In Chapter 3 the proposed approach is introduced followed by the experimental results in

Chapter 4. In the last Chapter a conclusion is made and further work to improve the proposed approach is suggested.

2 State of the art and examples of different Video Quality Metrics

In the following chapter we will go through a couple of different VQM's introduced. Also the category that each metric belongs to is mentioned based on the 3 different groups introduced in section 1.3.2 It should be mentioned that the metrics are grouped based on the group they should belong to although the authors claim the method belonging to some other group.

2.2 Full reference metrics

As mentioned in section 1.3.2.1 FR metrics are the metrics which have access to the reference video as well.

2.2.1 Mean Square Error (MSE)

The MSE method is based on comparing pixel by pixel differences in each frame and so it will find an overall value for each frame. In the first step the MSE value is calculated for each frame. In the second step the overall MSE value for the video is calculated based on the values calculated for each frame in the previous step. For example if we have a video consisting of T frames which each frame has a size of N by M pixels we will have:

$$\mathbf{MSE_t} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\left(x_{i,j,t}-y_{i,j,t}\right)^2}{NM} \qquad \textbf{Eq 1}$$

In Eq 1 $x_{i,j,t}$ is the pixel in the reference video which belongs to the coordinate (i,j) in frame t. Also $y_{i,j,t}$ is the pixel belonging to the coordinate (i,j) in frame t of the test video. After calculating $\mathrm{MSE_t}$ for each frame we will have to calculate Eq 2 for the overall quality of the video.

$$\mathbf{MSE} = \frac{\sum_{t=1}^{T}\mathbf{MSE_t}}{T} \qquad \textbf{Eq 2}$$

Although MSE is a simple FR metric and easy to implement with low calculation complexity but the overall result does not correlate well with the results achieved throw subjective test. This is because of different factors such as the fact that it does not take the content of the video into account. No matter what type of content we have in a video we will have the same approach regarding each pixel individually. It is a fact that not only in videos but even in images an individual pixel does not act as a single parameter and it is influenced by its surrounding pixels and the region it is placed in. This is why using methods such as MSE will sometimes give accurate results but other times will be giving inaccurate results and the accuracy of the results are depended to the content of the video.

2.2.2 Peak Signal to Noise Ratio (PSNR)

For calculating Peak signal to Noise Ratio (PSNR) which is also a FR metric we will have to first calculate the MSE value for the video. Eq 3 is used to calculate PSNR for a video.

$$\mathbf{PSNR} = \mathbf{10\log_{10}}\left(\frac{m}{\mathrm{MSE}}\right) \qquad \textbf{Eq 3}$$

In Eq 3 m is the maximum value that a pixel could take, for example in an 8 bit image the value of m will be 255. An important factor to keep in mind is that there is no agreement on how to apply the MSE or PSNR methods on color images or videos [**14**].

Since PSNR also faces each pixel separately all the criticism made towards MSE in section 2.2.1 could be also made for PSNR. With regards to all these criticism made towards the PSNR method and the fact that the correlation between its result and the Mean Opinion Score (MOS) are rather poor [**15**], Huynh-Thu et al [**16**] claim that PSNR will give good results when the video content and codec are fixed across the video. This is what we also mentioned previously that the overall score given to a video using this method is really depended to the content of the video with some videos having accurate results and some having a poor result.

2.2.3 Method proposed by Wang et al [17]

As most IQM and VQM's introduced by the Live research group [**18**] this method is also based on the SSIM metric [**19**] for IQA (such as [**17**], [**19**] and [**10**]). In this method instead of calculating the quality value for all 8 by 8 sliding windows in the frame they calculate the value for some randomly selected windows. The number of selected windows is not mentioned by the authors but they claim that a good number of sub-blocks will give robust results as well as reduce the calculation time. The quality measurement will be applied to the Y, Cb and Cr channels. After this step the overall value will be calculated throw adding up a weighted value of the values calculated for each channel. After calculating the overall quality of each frame these values will be added up to give an overall quality of the video. An important factor during this summation is that each frame is weight differently. Two factors that are used in weighting the overall quality are the motion of the frame and the darkness and brightness of the frame. The weighting is done in a way that dark regions are weight less than the bright regions. This is because during the observation process less time is put on the dark points. Also large motions will be weight less than a low amount of motion.

One of the problems that this FR method might face is the fact that selecting the sub-blocks in each frame on a random order might influence the overall result. In other words the result will be dependent on the content of the video, type of distortion and the level of distortion; also if the distortion is spread around the whole frame in a slightly equal manner or it's just applied to a specific part of the frame. Obviously if the distortion is just in a specific part of the frame we will be facing inaccurate results no matter if we take this region into account or not. If the region is among our sub-blocks we will rank the video quality lower than it is and if the region is not among our sub-blocks the video quality will be ranked higher than it should be. An example of this problem could be seen in Figure 5, in Figure 5 (a) the distortion could be seen on the upper half of the cameraman image while in Figure 5 (b) the distortion has been spread around the entire image. The problem we might face would be that in the case of Figure 5 (a) we will have different results depending on where we select our sub-blocks, if the sub-blocks are focused more in the upper half we will have a score which will rank the image in a lower quality than it should have and if we have more sub-blocks in the lower half we would have a score which will rank the quality in a better place than it should have. On the other hand in the case of Figure 5 (b) no matter where the sub-blocks are selected we will approximately have the same score for the whole image.

Also treating the luminance channel and the chrominance channels in the same way and with the same IQM does not seem to be the best option. Later on in Chapter 3 we will show that the amount and structure of information in these channels are totally different and we are facing totally different information in luminance channel compared to the chrominance channels.

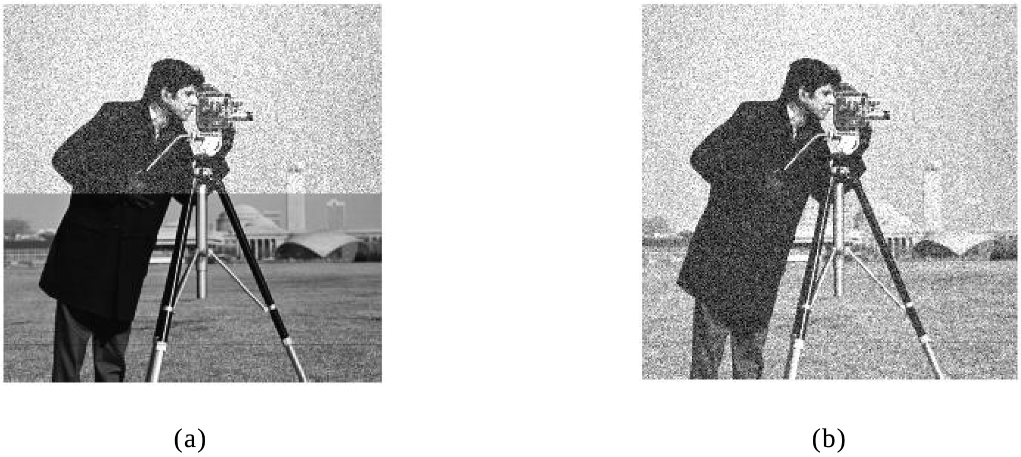


Figure 5. Example of an image (a) with distortion on a particular region, (b) with distortion spread around all the image.

2.2.4 Method proposed by Zhang et al [20]

This method is a FR VQM based on Singular Value Decomposition (SVD) of a complex matrix. They take the luminance channel of the frame as the real part of a new complex number and the chrominance channel of the frame as the imaginary part of the complex number. This selection is made because HVS is more sensitive to distortions in the luminance channel than the distortions in the chrominance channels. With this order we will have a new matrix presented in Eq 4.

$\mathbf{A(x,y) = Y(x,y) + C(x,y)i}$ **Eq 4**

In this equation, $\mathrm{Y(x,y)}$ is the luminance matrix and $\mathrm{C(x,y)}$ is the chrominance matrix. In color images in the format of YCbCr the luminance would be Y and the chrominance could be either one of Cb or Cr. In the next step $\mathrm{A(x,y)}$ which is the new matrix, is divided to blocks of 8 by 8 pixels. Using the method introduced in [21], SVD is applied on sub-blocks of the image. In the next step the quality of each frame is calculated according to the method introduced by Eskicioglu et al [22]. After this step we would have a score for each sub-block of each frame named $\mathrm{D_i}$. According to different researches made [23] and [24], photoreceptors are distributed unequally on the retina and this would affect the overall quality of the frame. To overcome this problem a weighting factor has been introduced named as $\mathrm{e_i}$ for each sub-block. With respect to the mentioned factors the overall quality of each sub-block is calculated in Eq 5.

$\boldsymbol{D_i(new) = D_i e_i}$ **Eq 5**

After calculating Eq 5, the overall quality of each frame ($\mathrm{CMSVD_f}$) is calculated in the same way as it was introduced in [22] for images. In the last step the average value of $\mathrm{CMSVD_f}$ of all frames is assigned as the overall quality of the video. In this method the maximum value of $\mathrm{CMSVD_f}$ is mentioned as well. This is because if a burst-of-error exits in a video the observer will rate the video quality lower than the average value [17]. Burst-of-error occurs when all the frames of a video except a few have high quality and those few have really low quality.

Although the use of SVD on a complex matrix constructed of the luminance channel and a chrominance channel seems really interesting especially that the method proposed by Eskicioglu et al

[**22**] is a rather famous metric for grey scale images, but on the other hand the fact that only one of the chrominance channels have been taken into account could be a drawback of this method. On the other hand it seems that even with only one chrominance channel taken into account the results are comparable with some of the state of the art metrics.

2.2.5 Method proposed by Lin et al [25]

This FR method focuses on low bit-rate video communication systems. The method is a modified version of the MSSIM method introduced in [**19**]. The modification tries to detect the global and local random spatial perceptual degradation that could be seen in low bit-rate videos. SSIM [**19**] is an objective image quality metric. On the other hand, MSSIM is the mean value of SSIM for each local window in the image. The steps taken in this method to assess the video quality in the modification process are as followed:

1. Video frame is divided into sub-blocks.
2. For each sub-block MSSIM is calculated.
3. The following values are calculated for each frame.

$$\mathbf{mnMSSIM} = \frac{\mathbf{1}}{\mathbf{I}} \sum_{\mathbf{i} \in \mathbf{I}} \mathbf{MSSIM_i} \qquad \textbf{Eq 6}$$

$$\mathbf{miMSSIM} = \mathbf{min}\{\mathbf{MSSIM_i}\}, \forall \mathbf{i} \in \mathbf{I} \qquad \textbf{Eq 7}$$

$$\mathbf{hierMSSIM} = [\mathbf{mnMSSIM}]^{\propto} \cdot [\mathbf{miMSSIM}]^{\boldsymbol{\beta}} \qquad \textbf{Eq 8}$$

In Eq 6, Eq 7 and Eq 8 i is the sub-block number and I is the total number of blocks. mnMSSIM is used to measure global perceptual distortions in the frame. miMSSIM is used to isolate local perceptual distortions. hierMSSIM is used to combine the two previous distortions calculated and give an overall value for the quality of the image. α and β which are positive values are selected according to the importance each factor has for us. In this work the values are set to one.

Although having a local as well as a global view on the quality seems to be a good approach to take but giving both the local and global factor the same weighting value does not seem to be the best choice. Obviously selecting the weighting factors with more precision will improve the accuracy of the results.

2.2.6 Method proposed by Pahalawatta et al [3]

In the proposed method, motion vectors are used to calculate temporal consistency metrics. These metrics are weighed and then converted in a single value which will present the overall temporal inconsistencies. The temporal consistency metrics used in this method are:

1. Motion vector consistency metric.
2. Motion estimated temporal difference metric.
3. Motion estimated temporal variation metric.

Although unlike most other metrics this FR VQAM has a temporal approach for evaluating the video quality but on the other hand not taking the spatial distortions into account is a drawback for the proposed method.

2.2.7 Method proposed by Ong et al [26]

In this FR method 5 different factors which will affect the visibility of distortion have been taken into account, luminance masking, texture masking, temporal masking, block fidelity and content richness fidelity. Normally the first 3 factors are taken into account when observing video quality but the authors here have added the last 2 as well which has improved the results compared to metrics that don't pay attention to such factors.

2.2.8 Method proposed by Kiemel et al [27]

This method tries to improve the results calculated from the PSNR method with a help from temporal pooling methods. We will first calculate the PSNR value for the Luminance (Y) channel which is then named as $\mathrm{PSNR}^{\mathrm{Y}}_{\mathrm{Mean}}$. But this is not the only value they calculate for the video. Other values are $PSNR^{Y}_{Min}$, $PSNR^{Y}_{Max}$, $PSNR^{Y}_{sDev}$, $PSNR^{Y}_{90}$ and $PSNR^{Y}_{10}$.they will also calculate Eq 9 for each frame which i is representing the frame number.

$$\boldsymbol{dPSNR^{Y}_{i} = \left|PSNR^{Y}_{i} - PSNR^{Y}_{i-1}\right|} \qquad \textbf{Eq 9}$$

They would then calculate the previous 6 values for $\mathrm{PSNR}^{\mathrm{Y}}$ and would have 12 values in total. In the next step these 12 values will be also calculated for the U and V channel ending in a total of 36 values. The method increases the correlation between objective and subjective test by 10% when it is compared to the normal PSNR method. Although the correlation rate is not even close to the correlation rate given by other methods introduced but the simplicity of the method is an advantage of this FR method.

2.2.9 Method proposed by Seshadrinathan et al [28]

This new FR method which has just been published in early 2010 is among the best metrics proposed till today. This metric which has been named MOVIE (MOtion-based Video Integrity Evaluation) has a temporal as well as a spatial approach giving the chance for the user to have three different values assigned to the quality of a video, a spatial quality value (Spatial MOVIE), a temporal quality value (Temporal MOVIE) and a spatial-temporal value (MOVIE) which is the result of multiplying the previous two values.

Before any processing on the videos both the reference and the test video are decompressed into a spatial-temporal bandpass channel using a Gabor filter. As most VQM's introduced by this research group the idea behind the spatial approach is inspired by the SSIM method. In the case of temporal approach the closer the MV of the test video to the corresponding MV of the reference video the higher the Temporal MOVIE value would be.

The fact that we are able to have three different values for each video is a great progress made. Also having the option of having a spatial-temporal approach has ranked this method among the best methods available so far.

2.3 Reduced reference metrics

As mentioned in section 1.3.2.2 RR metrics do not have the original reference video but there is some information regarding the reference video. It should be mentioned that the proposed approach we will have which is introduced in chapter 3 is a RR metric as well.

2.3.1 Method proposed by Albonico et al [29]

This method [**29**] is based on the VSSIM method [**17**] which was introduced in section 2.2.3. This method uses the same approach and formulas but instead of evaluating the video quality based on the reference video itself they use some information such as the mean and variance value. They also use a 16×16 sliding window and instead of calculating the covariance between the reference and test video they use an approximation of the value using the mean and absolute values of both videos along with a term that estimates the channel induced distortion. Because of all these approximation they also introduce a value which could be seen in a lookup table they provide with each video sequence, this value depends more on the quantization step.

Although this method is ranked among the RR methods but it is more or less the same VSSIM method [**17**] which is a FR metric. The difference is that for example in the case of video transmission, in this method all the values for the reference video is calculated in advance before transmission but VSSIM calculates these values at the receivers end. Although doing this would decrease data size but on the other hand all these estimations and the lookup table that should be calculated and provided in advance for every video sequence would decrease the accuracy of the results compared to the VSSIM method.

2.3.2 Method proposed by Fu-zheng et al [30]

This method is based on the digital watermarking technology. The procedure is made of 3 main steps:

1. Watermarking embedding.
2. Watermarking extraction.
3. Video quality assessment.

In summery in the procedure they use a known data as the watermark. After transferring the video, in the destination the quality of the watermark is assessed. Since HVS is less sensitive to luminance variation than chrominance variation the watermark has been embedded in the luminance component of the video sequence.

To reduce the effect of watermarking on the frame quality a watermark with the size 44 by 36 has been divided to n blocks. Each block is then embedded in every other frame of the video sequence. This would result the period of the watermark to be 2n. The place of each block in the watermark frame is determined by the sequence number.

To measure the quality of the watermark they introduced a new metric called Pixel Recovery Rate (PRR). PRR is calculated as shown in Eq 10.

$$\mathbf{PRR} = \frac{\mathbf{N_1}}{\mathbf{N}} \qquad \textbf{Eq 10}$$

In Eq 10, N_1 is the number of correctly retrieved pixels and N is the total number of pixels in the watermark.

Although the original watermark data should be available to calculate the PRR metric and therefore the method should be ranked as an RR metric, Fu-zheng et al rank their method among the NR metrics.

Regarding the method introduced, the fact that no evaluation is made on the chrominance channels is a drawback to the method. Although as mentioned before HVS is more sensitive to the luminance channel but still the chrominance channels have their effect on the video quality. Also the fact that the watermark will not cover the whole luminance channel and will just be inserted in some regions could influence the results in different ways. The same criticism made regarding the selection of random sub-blocks in section 2.2.3 could be also made for this method as well.

2.3.3 Method proposed by Farias et al [31]

This method like the one presented before [**30**], uses the watermarking approach to calculate the video quality. The difference between this method and the previous method [**30**] is that there has been some psychophysical experiments made before the proposal of the method. These experiments have been made to measure the visibility threshold and the mid-annoyance values. Therefore there is more focus on watermarking than the previous work [**30**]. For this reason and after the psychophysical tests the strength of the watermark is different, depending on different videos.

Like the previous method [**30**] this method has also been grouped in the RR metric. Most of the criticisms made in section 2.3.2 regarding the method introduced in [**30**] could be also made for this method as well.

2.4 No Reference metrics

As mentioned in section 1.3.2.3 NR metrics are metrics which there is no information regarding the reference video. Therefore the metric will try to evaluate the amount of distortion in the video and evaluate the quality based on the calculated value.

2.4.1 Method proposed by Yang et al [9]

Unlike most VQA methods that have a spatial approach this NR method has a temporal approach for evaluating the quality of a video. HVS is sensitive to sharp changes and so consecutive frames in a video tend to have quite the same structure. Keeping that in mind, in the first step the authors calculate the Motion Vectors (MV) for each pixel of a frame. In the next step they find the same regions in consecutive frames according to the motion vectors each pixel has in that region. This is done by selecting a threshold range so that if the MV of a specific pixel is in the range of a specific region the pixel belongs to that region otherwise they would try to find the corresponding region for that pixel. In the last step each matching region is compared to give an overall quality of that region. By calculating the quality of all the regions in a frame we would have a score for each frame and later on for the whole video sequence. The metric they use is a differential approach towards the regions. First the difference between a region and the same region in the previous frame is calculated and then a Gaussian filter is applied to the frame and the difference is calculated again to have 2 different difference values for each region.

The proposed approach in this master thesis came from the idea given in this method but with respect to some criticisms made. Calculating the MV’s for each pixel is a time consuming work and will take a huge time especially in videos with high frame rates. Also we will show in section 3.2.1.1 that using a Mean Absolute Difference (MAD) is not the best method for calculating the motion vectors especially when we are trying to evaluate the quality of a video. In this case every small error will add up and therefore influence the final result in a huge way. Using the MAD method and then calculating the quality based on a difference method is using the same method for image quality twice. Although using a

Gaussian filter would help us in removing distortions but on the other hand all these distortions are affecting the quality of the video itself and trying to remove them from the video does not seem to be a good idea.

2.4.2 Method proposed by Opera et al [32]

In this method first a visual attention model is applied on the frames. In the next step quality of the regions which according to the model seems to be more significant are evaluated. Although according to the authors any kind of quality metric can be used but a simple blurriness metric is applied in this method. The visual attention model used is the model introduced in [**33**]. In summery the method mostly focuses on the visual attention model rather than the quality assessment part.

The fact that the regions which are not calculated to be in the region of interest in this NR metric are not evaluated in the quality evaluation of the video is a huge drawback. Although there should be more attention paid to the regions which will take more interest of the observer compared to the regions with lower importance but omitting the other regions is not a good idea. In our proposed approach to overcome this problem we have weighted each region depending on their importance for an observer.

2.4.3 Method proposed by Larabi et al [34]

In this NR metric the authors have taken a 3D multispectral wavelet transform approach to estimate the quality of colored videos. By doing this they will decompose the color video sequence into a number of spatial-temporal frequency channels each responding to a certain spatial-temporal location. After applying a 3D wavelet decomposition on the video they apply a spatial-temporal CSF filtering and also a luminance sensitivity masking to find the most attractive wavelet coefficient seen by the observer and in the next step an activity masking is done. And finally all the data is pooled.

The fact that the proposed method is trying to model some of the fundamental properties of the HVS is a great progress made by this metric compared to previous metrics. Keeping that in mind and with the approach taken the results show a good correlation with the subjective results.

3 Proposed approach

3.1 Introduction

With respect to the previous proposed metrics there still is a need for a metric which has a spatial-temporal approach and is based on QOE. Figure 6 is a schematic of what the first idea behind the proposed metric was. The video is processed in the spatial and also temporal domain. There is information regarding QOE extracted from the video and sent to the pooling system as a controlling factor and there is also a link between the spatial domain and the temporal domain.

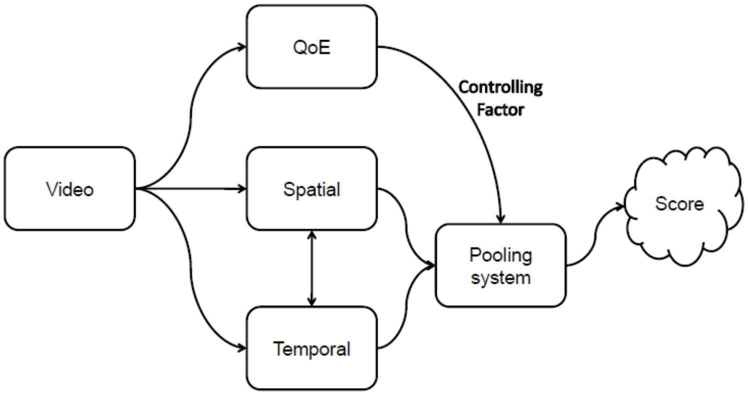


Figure 6. A schematic of what the proposed method is based on.

The method introduced in this work is a "Reduced Reference Video Quality metric". As it can be seen in Figure 7 apart from the test video that will be used as an input for the metric, we will also use the motion vectors of the reference video along with the motion activity density and the overall quality of the reference video which later on will be used as the information we have on the reference video.

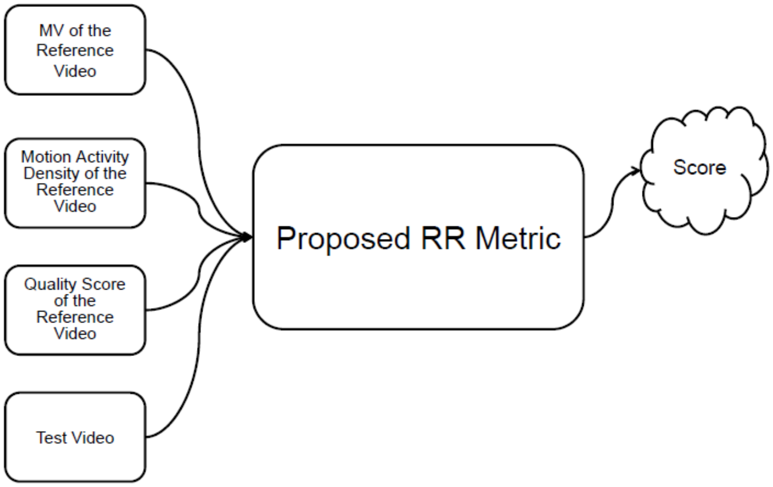

MV of the Reference Video
Motion Activity Density of the Reference Video
Score
Proposed RR Metric
Quality Score of the Reference Video
Test Video

Figure 7. Structure of the proposed quality metric.

Figure 8 gives an idea of what is the structure inside the “Proposed RR Metric” block in Figure 7, grey shaded blocks are used in the “proposed RR Metric” block. As it can be seen if we add the “Motion Activity Density of the Reference Video” block the metric would be what we introduced in Figure 6.

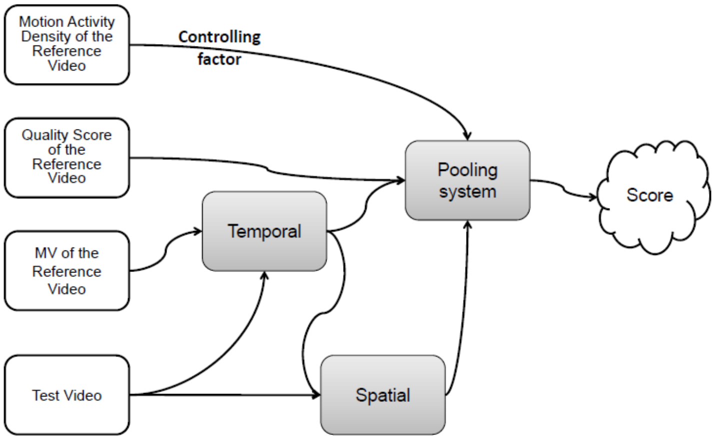

Motion Activity Density of the Reference Video
Controlling factor
Quality Score of the Reference Video
MV of the Reference Video
Test Video
Temporal
Spatial
Pooling system
Score

Figure 8. An overall view of how different parts of the proposed metric are in connection with each other.

The proposed method is based on the fact that the HVS is sensitive to sharp changes in the video sequence. In other words in order to evaluate a video, the video stream would need to be in a continuous format and we cannot have consecutive frames with totally different structures. In the method we have taken advantage of this simple fact. In our method, we divide each frame (for example frame *i*) to different sub-blocks and try to find the matching sub-block in the next frame (in this case in frame *i+1*) and then a FR IQM is used to estimate the quality of the matching sub-blocks in a frame. In the next step we will calculate the overall quality of each channel and at the last step we will pool the result so that the quality of the video is measured. The following factors are taken into account and used in the pooling step:

1. Quality score of the Y channel.
2. Quality score of the U channel.
3. Quality score of the V channel.
4. Number of matching motion vectors between the reference and the test video.
5. Quality score of the reference video.
6. Motion activity density of the reference video.

An important factor to keep in mind is that to have results closer to the results given by human observers we have also used a Visual Attention Model (VAM) [**35**] to weight our results during the quality measurement of the common regions.

3.2 Temporal approach

Figure 9 shows an overview of different parts of the temporal approach that we will take. As it can be seen we will first calculate the MV’s or in other words find the matching sub-blocks in consecutive frames. We will then compare the MV’s of the reference video with the MV’s of the test video. This information will be sent to the pooling block to be pooled in the last step. We will also select the correct matching sub-blocks and use them when we are calculating the quality of the luminance and chrominance channels in the spatial approach which we will talk about later in the method.

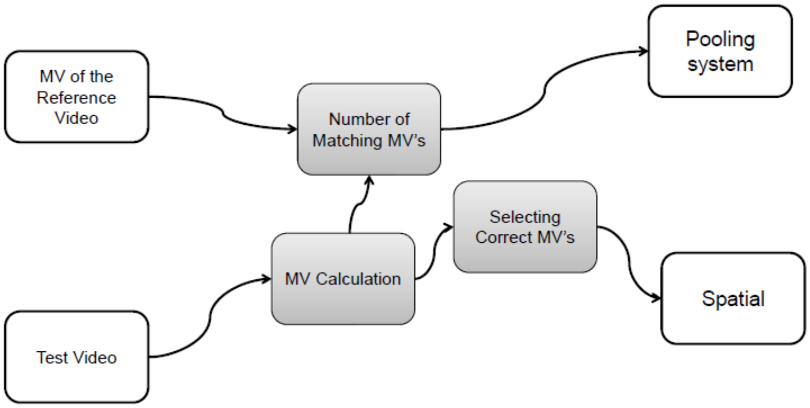


Figure 9. An overview of the temporal approach.

3.2.1 Calculating the common regions

For finding the common regions between a frame and the next frame we should use a block matching algorithm. Barjatya [**36**] has made a good review on 7 different block matching algorithms and in conclusion it shows that the Adaptive Rood Pattern Search (ARPS) method [**37**] is the best method among the block matching algorithms both in the case of calculation time and also correct results. This is why we used the APRS method for calculating the common regions. It should be mentioned that the common regions are only calculated for the luminance channel and then the coordinates are used to apply IQM's on the chrominance channels as well.

By calculating common regions we are also calculating the MV's for each frame as well, Figure 10 is an example of the procedure taken. For frame i we will take a sub-block of 16×16 and search for the same sub-block in frame $i + 1$.

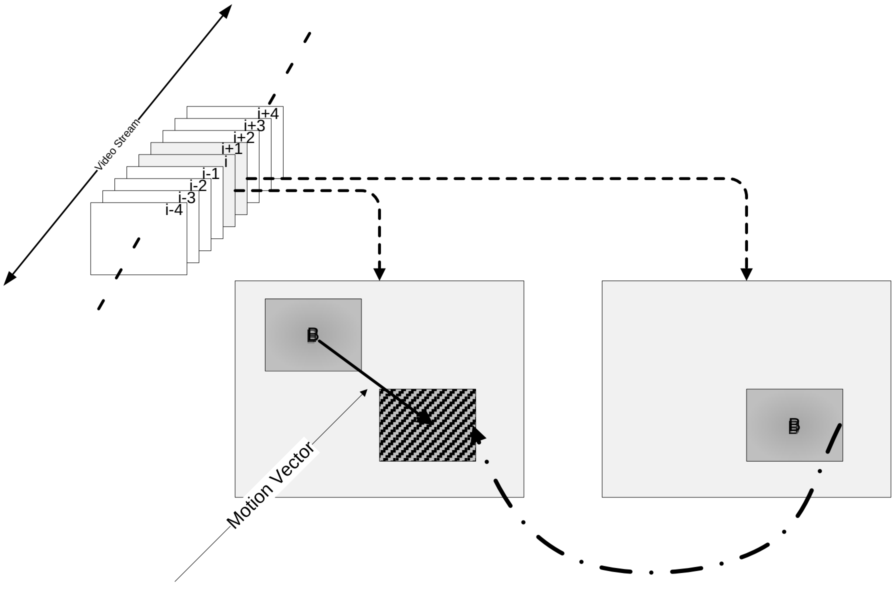


Figure 10. Finding the common regions and hence calculating the Motion Vectors.

3.2.1.1 ARPS block matching algorithm [37]

A common expectation in a video sequence is that there is a good chance that the neighboring sub-blocks would have approximately MV's with the same direction and magnitude. In their proposed method Nie et al [**37**] have taken advantage of this fact and so, in order to calculate the MV for a sub-block, they use the information available from the MV of the sub-block which is on the left of the current sub-block. In the first step the algorithm assumes that for a specific sub-block the MV is equal to the MV of the sub-block on the left. Then the pointed sub-block along with the rood pattern distributed points are checked, as shown in Figure 11. The step size taken in this point is also based on the predicted motion vector. Eq 11 shows how the step size is calculated.

$\boldsymbol{S} = \mathbf{max} \quad (|\boldsymbol{X}|, |\boldsymbol{Y}|)$ **Eq 11**

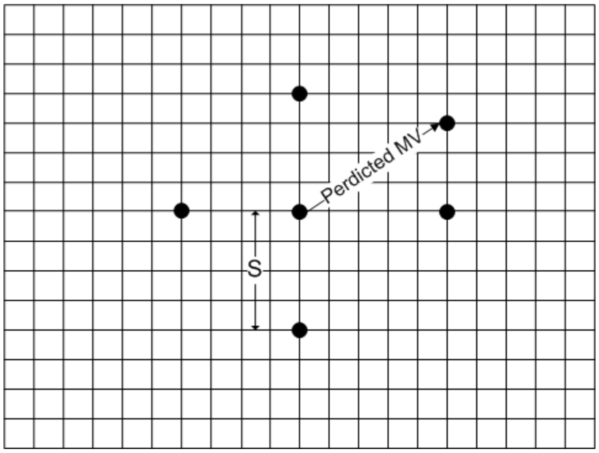


Figure 11. Points searched in the first step using the ARPS method.

In Eq 11, X and Y are the coordinates of the predicted motion vector. After this first step the sub-block which has the most possibility to be our corresponding sub-block is selected. In the next step we would start a Small Diamond Search Pattern (SDSP) and continue this procedure until we find the best matching sub-block.

Although most, if not all block matching algorithms use a Mean Absolute Difference (MAD) approach to calculate the best matching sub-block, we use the Complex Wavelet Structural Similarity (CW-SSIM) IQM [**38**] to find the best suitable sub-block. This is because when the MAD method is used, we are using a FR IQM which is acting like a simple difference metric such as the MSE method. In this way we would be using IQM's twice, first during finding the matching block and the second time when we want to calculate the quality of each frame. Also using CW-SSIM will help us in calculating better and more robust results compared to when using the MAD method. Table 2 shows the MSE value for the difference between the original frame and the reconstructed frame using the two different methods. As it can be seen using CW-SSIM will lead us to better results. Figure 12 shows the difference between the original and the reconstructed frame using the CW-SSIM method. Also in Figure 13 we can see the difference between the original frame and a reconstructed frame using the MAD method. The frame used is from one of the videos of the "LIVE Video Quality Database" [**39**], [**40**] and [**41**] which we will later on test our proposed method based on it as well.

Table 2. MSE values between the original frame and the reconstructed frame using CW-SSIM and MAD methods

Method used for reconstruction of the frame	MSE value between the reconstructed frame and the original frame
CW-SSIM	29.3877
MAD	37.22332

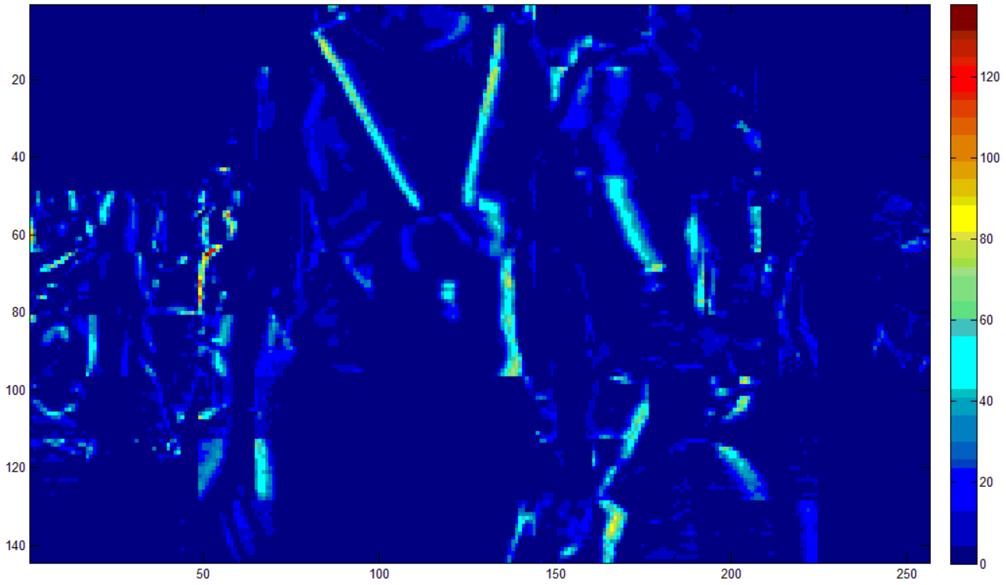


Figure 12. Difference between the original and the reconstructed frame using the MV's calculated with the CW-SSIM method.

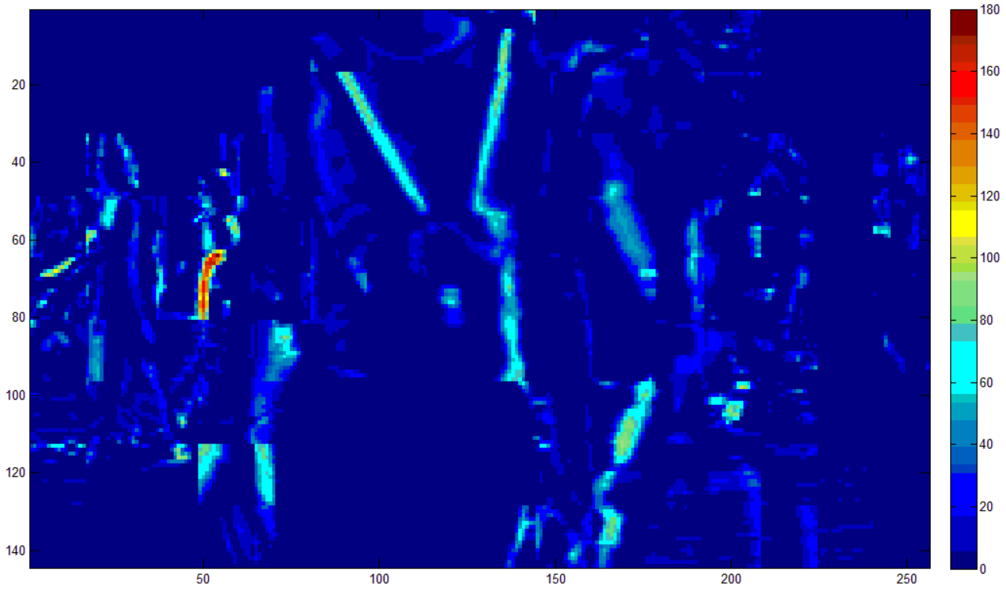


Figure 13. Difference between the original and the reconstructed frame using the MV's calculated with the MAD method.

3.2.1.2 Complex Wavelet Structural Similarity (CW-SSIM) image quality metric [38]

CW-SSIM is a FR IQM which is based on the SSIM method [**19**] but this time the method uses complex wavelet to evaluate the quality of the image or as the authors mention in [**38**] *"the CW-SSIM index is an extension of the SSIM method to the complex wavelet domain"*. The most important reason for using the CW-SSIM method is that when the CW-SSIM index is calculated in the complex wavelet domain the metric will be no longer sensitive to nonstructural geometrical image distortions. When working with videos, we will face changes such as rotation, scaling and etc. most IQM's are sensitive to these changes but an improvement in CW-SSIM compared to previous methods is the insensitivity of it towards these changes and these changes are what we will face in videos when comparing two consecutive frames.

CW-SSIM is calculated using Eq 12. In this equation we have $c_x = \{c_{x,i}|i = 1, \dots, N\}$ and $c_y = \{c_{y,i}|i = 1, \dots, N\}$ which are two sets of coefficients extracted at the same spatial location in the same wavelet sub-bands of the two images being compressed. The complex conjugate of c is c^* and K is a small positive constant mainly used to improve the results when Signal to Noise Ratio (SNR) is low.

$$\tilde{S}(c_x, c_y) = \frac{2\left|\sum_{i=1}^{N} c_{x,i} c_{y,i}^*\right| + K}{\sum_{i=1}^{N} |c_{x,i}|^2 + \sum_{i=1}^{N} |c_{y,i}|^2 + K} \qquad \textbf{Eq 12}$$

Like SSIM, the maximum value of CW-SSIM occurs when the two images compared are identical and in this case the value of CW-SSIM is 1.

3.2.2 Number of matching motion vectors

After calculating the motion vectors for the reference and the test video we will calculate one of the parameters in the pooling procedure, "Number of matching motion vectors". This parameter is calculated easily just by comparing the motion vectors for the two videos. Obviously in order to have a video which matches the quality of the reference video our motion vectors should be the same and any difference between the two sets of MV's will affect the quality of the videos as well. To calculate this value we will check if $MV_{R(i,j)}$ which is the j^{th} motion vector in frame i in the reference video is equal to $MV_{T(i,j)}$ which is the j^{th} motion vector in frame i in the test video. After calculating the number of matching MV's, in the last step we will divide the resulted value by the number of MV's in the video and assign a value for each video named as $NMMV$ which will show the percentage of the matching motion vectors between the reference video and the test video.

3.2.3 Selecting correct motion vectors

After calculating motion vectors for a frame, we should check that they do really help us in finding the common regions in the frame. This is because no matter a specific region in one frame is in the next frame or not the block matching algorithm will find the closest region in the next frame corresponding to that region. The problem arises when a region which is not corresponding to the same region in the previous frame is selected and measurements are done between these two regions.

It is assumed that in a video, the MV's will more or less have the same direction and magnitude. With respect to this, we will have the corresponding motion vectors $MV = \{MV_i | i = 1, \dots, N\}$ for each frame with N representing the number of motion vectors in a frame. We will then calculate Eq 13 which in this equation, m is the mean value of all motion vectors in that specific frame, this way we will find which direction and magnitude the motion vectors will probably have. In the last step a threshold T is selected and if $|MV_i| < Tm$ we will take the region corresponding to MV_i into account otherwise we will skip this region.

$$\boldsymbol{m = \frac{\sum_{i=1}^{N} MV_i}{N}} \qquad \textbf{Eq 13}$$

Figure 14 shows an example of what MV's in a frame would look like. In Figure 14 MV's which have a dashed pattern will not be accepted because they are out of our supposed threshold and therefore the grey sub-blocks which correspond to these MV's would not be taken into account as well.

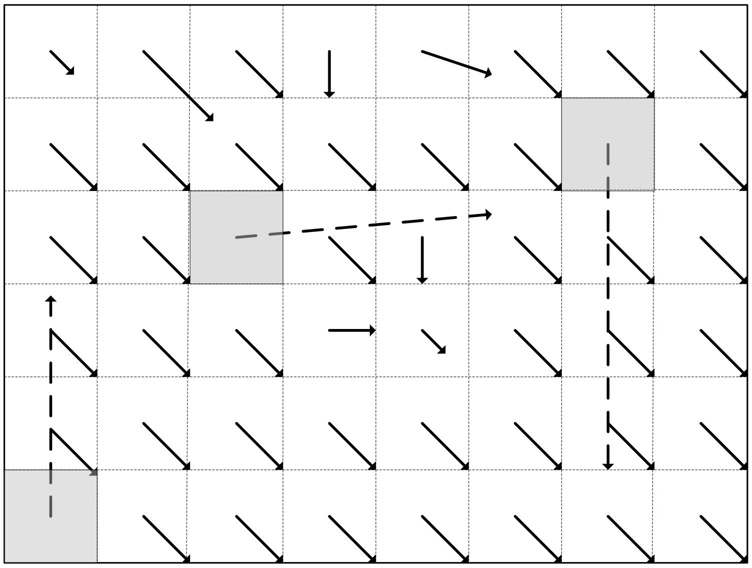


Figure 14. Example of the MV's in a frame.

3.3 Spatial approach

Figure 15 shows the procedure taken during the spatial approach of the proposed metric. As it can be seen we will first use a visual attention map to estimate the importance of each sub-block so that we can use this value as a weighting factor for sub-blocks in the frames. Using this information along with the information gathered from the temporal approach will help us in calculating the quality of different channels in each frame. After calculating the quality of the channels we will use these values in the pooling system.

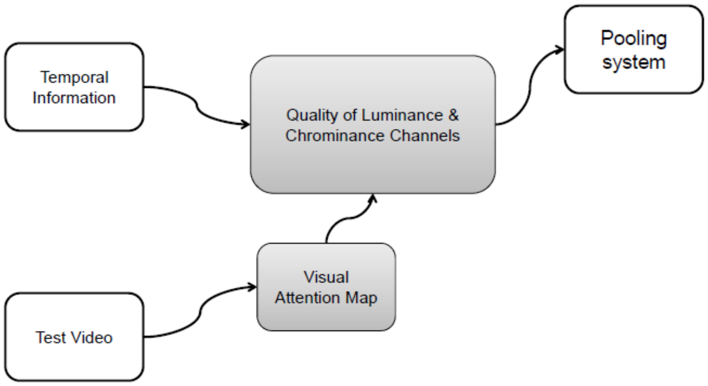


Figure 15. An overview of spatial approach, grey shaded blocks are the procedures taken.

3.3.1 Visual attention map

With the speed that the frames are being played in a video it is an obvious fact that we are not able to pay attention to all the details available in a single frame. In other words the observer would have enough time to pay attention to regions that would somehow stick out of the video and make him/her pay more attention to them. Since in image/video quality assessment models we are trying to follow the HVS and have results that are close to the subjective results, finding these salient regions in the frames and giving a higher value to them compared to other regions would increase the accuracy of the metric. This is why applying a saliency map to the frames and weighting each region based on the attention map would increase the accuracy of the objective results.

Achanta et al [**35**] introduced a method for capturing salient regions in colored images. As well as being a simple and easy to implement method, it also gives correct and robust results which outperforms 5 previous state of the art metrics [**42**], [**43**], [**44**], [**45**] and [**46**]. The method could be summarized in Eq 14, which $s(x, y)$ is the saliency score for each pixel (x, y), I_μ is the mean image feature vector and $I_{\omega hc}(x, y)$ is the corresponding image pixel vector value in the Gaussian blurred version of the reference image, here we use a 5×5 separable binomial kernel. The saliency map is calculated in the Lab colorspace and $\| \ \|$ is the Euclidean distance.

$$\boldsymbol{S}(\boldsymbol{x}, \boldsymbol{y}) = \left\| \boldsymbol{I}_{\boldsymbol{\mu}} - \boldsymbol{I}_{\boldsymbol{\omega hc}}(\boldsymbol{x}, \boldsymbol{y}) \right\| \qquad \textbf{Eq 14}$$

Using this method in our metrics there are a couple of main points to keep in mind regarding the use of attention maps in the proposed method:

1. Since this metric is a metric used in images and not videos we applied this metric to each frame of our test video.

2. We normalized the output result for each frame in a way that the highest value in the attention map will be 1. In other words we divided the value for each pixel by the maximum value assigned to a pixel in that frame so that the most important pixel will have a value of one assigned to it.
3. Although this method will give different values for each pixel but since we are working with sub-blocks in the frame we will calculate the mean value of all the pixels in each sub-block and assign the resulted value to that sub-block representing the importance of that specific sub-block.
4. We only use the results from running this method on the luminance channel. This is because that there is not really much information available in the chrominance channels. Figure 16 shows a frame from a video sequence. As it can be seen in Figure 17 which shows the Y channel of the same frame there is a lot of information in the luminance channel. On the other hand as Figure 18 and Figure 19 show there is not much information in the Chrominance (U and V) channels of the frame.

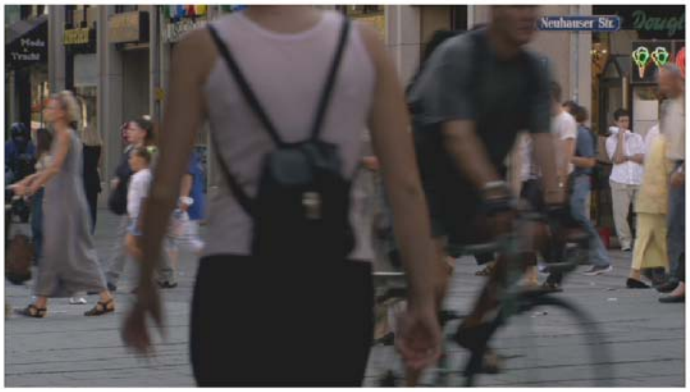


Figure 16. Original frame of a video sequence.

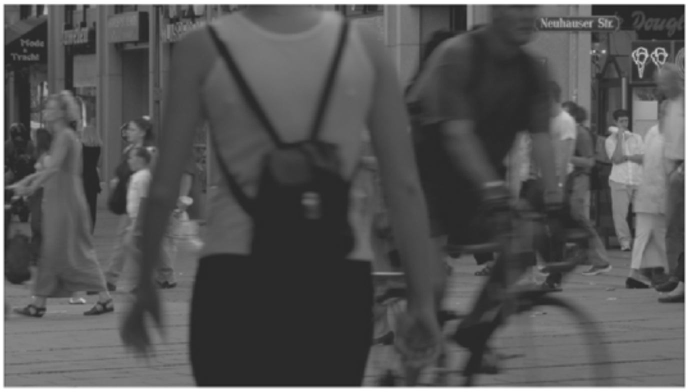


Figure 17. Luminance channel of the frame shown in Figure 16

Figure 18. U channel of the frame shown in Figure 16

Figure 19. V channel of the frame shown in Figure 16

3.3.2 Calculating the quality of the frames

For calculating the quality of the frames we apply IQM's to each frame separately. Furthermore we will approach the evaluation of the channels in two different ways. We would have one method for the luminance channel and would take another approach for the chrominance channels. An important step in calculating the quality of the channels is that we will only find the quality of the sub-blocks which correspond to correct MV's that we calculated in section 3.2.3.

3.3.2.1 Calculating the quality of the luminance channel of frames

For calculating the quality of the luminance channel we use the CW-SSIM method [**38**] which was introduced in section 3.2.1.2.Each sub-block in frame i will be compared with the matching subblock in frame $i+1$ using the CW-SSIM method. For example if we have N different correct subblocks in frame i for each subblock we will have:

$$\boldsymbol{QSY(i,j) = CW - SSIM\big(sbY(i,j), sbY(i+1,j)\big)} \qquad \textbf{Eq 15}$$

In Eq 15, $QSY(i,j)$ is the quality score of sub-block j in frame i. $sbY(i,j)$ is the sub-block j in frame i and $sbY(i+1,j)$ is the sub-block in frame $i+1$ that matches with sub-block j in frame i. If $AMS(i,j)$ would be the visual attention map score for subblock j in frame i then we would have:

$$\boldsymbol{SY(i,j) = QSY(i,j) \times AMS(i,j)} \qquad \textbf{Eq 16}$$

Eq 16 will give us the overall score for sub-block j in frame i which is shown by $SY(i,j)$. Using Eq 17 will calculate the overall quality of a frame i which is simply a mean value of the sub-block scores.

$$\boldsymbol{FSY(i) = \frac{\sum_{j=1}^{N} SY(i,j)}{N}} \qquad \textbf{Eq 17}$$

To calculate the overall quality of the luminance channel in the video sequence we will use Eq 18. In this equation YS will be the overall quality of the luminance channel in the video sent to the pooling system and M is the total number of frames in the video.

$$\boldsymbol{YS = \frac{\sum_{i=1}^{M} FSY(i)}{M}} \qquad \textbf{Eq 18}$$

3.3.2.2 Calculating the overall quality of the chrominance channel of frames

As mentioned and shown in section 3.3.1 although we do not have a huge amount of information in the chrominance channels but neglecting this channel would not be a good decision to make. For this reason we decided to use a simple MSE metric for these channels. Using an MSE metric will also reduce the complexity of calculation as well. Eq 19, Eq 20 and Eq 21 show the steps that we should take to calculate the overall quality of the chrominance channels in the video sequence. Obviously these values should be calculated for each one of the chrominance channels individually. In the equations, $sbC(i,j)$ is the sub-block j in frame i and $sbC(i+1,j)$ is the sub-block in frame $i+1$ that matches with sub-lock j in frame i, also $QSY(i,j)$ is the quality score of sub-block j in frame i. $FSC(i)$ is the overall quality of the chrominance channel of the frame i and CS is the overall quality of the chrominance channel of the video. For instance in the case of having a video in the YUV color space we will have US and VS corresponding to quality score of the U and V channel of our video sent to the pooling system.

$$\boldsymbol{QSC(i,j) = MSE\big(sbC(i,j), sbC(i+1,j)\big)} \qquad \textbf{Eq 19}$$

$$\boldsymbol{FSC(i) = \frac{\sum_{j=1}^{N} QSC(i,j)}{N}} \qquad \textbf{Eq 20}$$

$$\boldsymbol{CS = \frac{\sum_{i=1}^{M} FSC(i)}{M}} \qquad \textbf{Eq 21}$$

3.4 calculating the motion activity density of a video

As mentioned in section 1.2 depending on the density of motion in a video the observer will react differently to the same type and level of noise. In order to have a VQM which is leaning more towards the QOE rather than QOS we should take the intensity of motion activity in the video into account.

With regards to motion activity we could classify a video of a news anchor reading the news as a video with low activity and a close up video of a 100m running competition as a video with high activity. Although different methods have been used for calculating the motion intensity [**47**], [**48**], [**49**], [**50**], [**51**], [**52**], [**53**], [**54**], [**55**], [**56**], [**57**] and [**58**] but most of them are mainly based on statistical approaches on MV's. These methods mainly try to find a relationship between the motion activity intensity in a video and one or more statistical feature of the MV's such as the mean, median, standard deviation, maximum and … values of the MV's.

For example in MPEG-7 motion descriptors the standard deviation of the magnitude of the MV's would help us in sorting the videos into 5 different motion activity groups. [**47**] Gives a good description of this method and from trying this method on the MPEG-7 dataset it has obtained Table 3 which the thresholds given could be used for sorting the videos.

Table 3. Thresholds used for evaluating the motion activity in a video [47].

Activity value	Range of σ (Std. Dev. Of motion vector magnitude)
1	$0 \leq \sigma < 3.9$
2	$3.9 \leq \sigma < 10.7$
3	$10.7 \leq \sigma < 17.1$
4	$17.1 \leq \sigma < 32$
5	$32 \leq \sigma$

Peker et al [**48**] present 2 different approaches using 9 different statistical values calculated from the MV's. These values are:

1. Average of motion vector magnitudes (avg).
2. Median of the motion vector magnitudes (med).
3. Variance of the motion vector magnitudes (var).
4. $mean0$ which is calculated by subtracting the average of motion vectors in a frame and then calculating the magnitude and at last averaging.
5. $mean1$ which is different from $mean0$ in the sense that the frame is divided by 12 blocks and the vector average is computed for each block separately.
6. Temporal differentiation of the motion vector field ($diff$).
7. Maximum of the motion vector magnitudes (max).
8. Maximum of the motion vector magnitudes disregarding the top 1.5% ($max1$).
9. Maximum of the motion vector magnitudes disregarding the top 10% ($max2$).

In the first approach they calculate the nine mentioned values for every P frame in the video and assign a zero value to all the intra-coded frames. In the next step they calculate the average of these values for all the frames in the video and so at the end we will have 9 values for each video. 5 different motion density groups have been introduced as the following:

1. Very high.
2. High.
3. Medium.
4. Low.
5. Very low.

The proposed method is tested against a ground truth data set [**59**] and Table 4 shows the average absolute difference between the quantized descriptors and the ground truth.

Table 4. Average error for the 9 mentioned descriptors [48**].**

	max1	max2	var	max0	mean0	mean1	avg	median	diff
Average error	0.730	0.743	0.746	0.746	0.781	0.792	0.816	0.824	0.826

In the second approach Peker et al used the 9 introduced descriptors to evaluate the activity density in one video against another video without assigning the videos into different groups. For this, they gave the following definition for using the subjective database they were using in the previous approach:

"*The activity level of clip i is greater than the activity level of clip j if and only if all the subjects assign a higher activity level to the clip i than they assigned to clip j.*"

Using this definition they applied the descriptors on 4134 different pairs of videos. Table 5 shows the results from testing the 9 descriptors on the 4134 pairs they had.

With respect to the second approach introduced in [**48**] we selected five different videos which each belonged to one of the different motion activity density groups. For determining the motion group a video belongs to, we selected a combination of *max2*, *var* and *median* as the descriptors which we will take into account and determine which group a video belongs to. We have given the first priority to *median* value then the *max2* value and then the value for *var*. This means that in the case that the median values were equal we will observe the *max2* values and if they were equal as well the value for *var* will decide which group a video belongs to. The motion group that each video belongs to will then be used as a controlling factor in the pooling system.

Table 5. Number and percentage of pairs which the descriptors fail [48**].**

	max1	max2	var	max0	mean0	mean1	avg	median	diff
Number of errors	494	218	318	827	318	389	416	337	501
Percentage (%)	11.9	5.3	7.7	20.0	7.7	9.4	10.1	8.2	12.1

3.5 Pooling the overall score

As mentioned in the previous sections we will now have six different values in our pooling system ready to be pooled. These values are:

1. Overall quality of the Y channel which will be calculated using the CW-SSIM method and also being weighted according to the calculated attention map for the corresponding frame.
2. Overall quality of the U channel which will be calculated using the MSE method.
3. Overall quality of the V channel which will be calculated using the MSE method.
4. Number of matching Motion Vectors between the reference video and the test video.
5. Motion activity density of the reference video. In other words which of the five motion groups does the video belong to?
6. Quality score of the reference video.

In the first step of the pooling procedure we will determine the pooling function we are going to use depending on which motion group the test video belongs to. In other words we will have five different pooling procedures, one for each motion group and will apply one of them to the test video depending on what motion group is assigned to the reference video of the test video we are working with. All these five different procedures would follow the same steps till the last step that they calculate the pooling function based on the learning data we have in each motion group.

3.5.1 Change the pattern of the quality score for the luminance channel

As it is know, when we use the MSE function the higher the output value the lower the quality. On the other hand in the CW-SSIM method keeping in mind that the value calculated would have a maximum of one, the higher the value the higher the quality. In the case of the attention maps since we normalized the attention map we would have a higher value (which would also be one in the normalized case) for the regions with higher priority. With these descriptions and in a video with a YUV colorspace we will have the following structure for quality values in different channels:

- In the Y channel we will have a value between zero to one. The higher the calculated value the closer the quality of the test video is compared to the reference video.
- In the U channel we will have a value greater than zero with no maximum limit. The lower the value the closer the test and reference video match with each other.
- In the V channel we will have values following the exact description given for the U channel.

With respect to the points mentioned above we will make a small change in the values calculated for the Y channel so that all three channels follow the same pattern. The change is that we will subtract the value calculated from the maximum possible value which is one, Eq 22 shows the procedure. This will result in the fact that the better the quality of the test video the lower the value will be. This will enable us to have the same pattern for all the values calculated for the three different channels.

$$\boldsymbol{YS_{new} = 1 - YS} \qquad \textbf{Eq 22}$$

3.5.2 Standardizing the calculated values for different video sequences

After the change made for the values of the three channels we will calculated the mean value (Eq 23) between the following four parameters:

1. YS_{new}
2. US
3. VS
4. $NMMV$

$$\boldsymbol{MeanS} = \frac{\boldsymbol{YS_{new}+US+VS+NMMV}}{\boldsymbol{4}} \qquad \textbf{Eq 23}$$

In Eq 23, $MeanS$ is the mean value of the four different parameters. Also the value assigned as the overall quality value of the reference video would be calculated from Eq 23 and the next step is only done for the test videos. In the last step before pooling the score, for calculating the overall quality value of the test videos we will subtract the mean value calculated for each test video with the mean value calculated from the reference video. In this way we will try to standardize the values for different video sequences as well. Eq 24 shows the procedure taken, $MeanS_{st}$ is the new standardized mean value, $MeanS_{Test}$ is the mean value scores for the test video and $MeanS_{Reference}$ is the mean value of the scores for the reference video.

$$\boldsymbol{MeanS_{st}} = \boldsymbol{MeanS_{Reference}} - \boldsymbol{MeanS_{Test}} \qquad \textbf{Eq 24}$$

The mean values before and after standardization for 8 different video sequences is shown in Table 6. Also a plot of mean values before the standardization and after standardization could be seen in Figure 20. As it can be seen the values are spread around between 0.5 to 2.5 when they are not normalized but after normalization we see the mean values focus more or less around one value and are between 0.02 and 0.73.

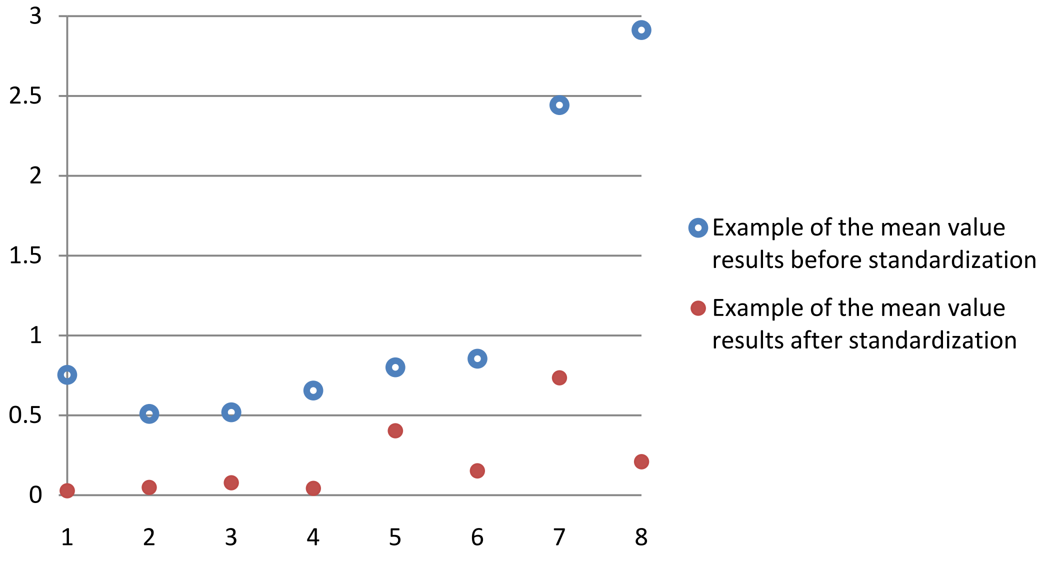


Figure 20. Example of the mean value results before and after standardization.

Table 6. Example of the mean value results before and after standardization.

Sequence number	$MeanS_{Test}$	$MeanS_{st}$
1	0.7531	0.0273
2	0.5089	0.0485
3	0.5187	0.0773
4	0.6550	0.0425
5	0.8003	0.4031
6	0.8542	0.1519
7	2.4427	0.7350
8	2.9130	0.2090

Having the $MeanS_{ST}$ value for all the videos in our database we will have to calculate our pooling structure based on the motion group a video belongs to. For each motion group we will use a neuro-fuzzy approach to train our system and then test our database on the resulted system. For this we use Matlab's fuzzy toolbox and its ANFIS (Adaptive Neuro-Fuzzy Inference System) structure. For training the algorithm we will use a hybrid optimization method with three two sided Gaussian membership functions (MF's) as its input and a linear output. The MF's used will be shown later in chapter 4 when we introduce the experimental results.

3.6 Overall view of the method

The overall schematic of the proposed method can be seen in Figure 21.

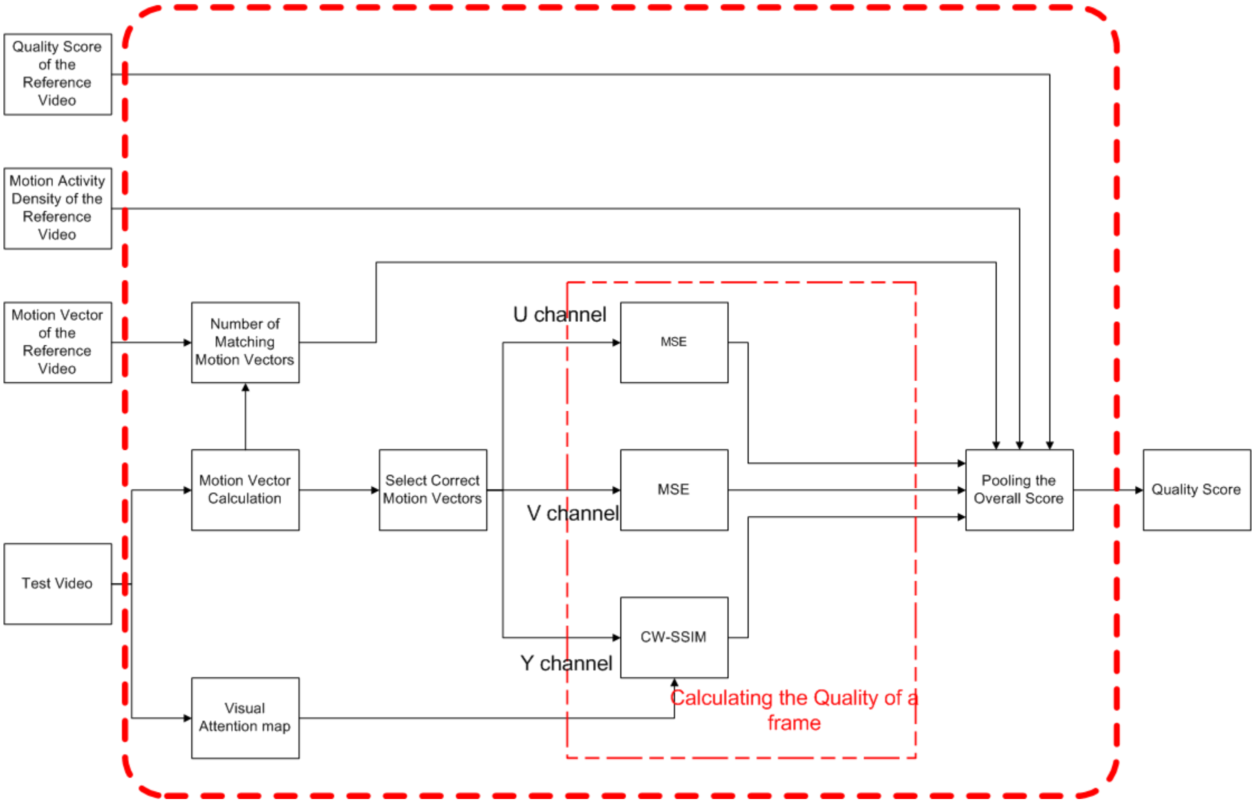

Quality Score of the Reference Video
Motion Activity Density of the Reference Video
Motion Vector of the Reference Video
Number of Matching Motion Vectors
Motion Vector Calculation
Select Correct Motion Vectors
Test Video
Visual Attention map
U channel
V channel
Y channel
MSE
MSE
CW-SSIM
Calculating the Quality of a frame
Pooling the Overall Score
Quality Score

Figure 21. schematic of the proposed method.

4 Experimental results

4.1 Introduction to the “LIVE Video Quality Database”

To test our proposed method we used the “LIVE Video Quality Database” [**39**], [**40**] and [**41**]. The database is consisted of videos with a spatial resolution of 768 × 432 pixels. There are two different frame rates in the database, 25 and 50 frames per second. There are also four different distortion and compression categories in the database which are:

- Wireless distortions
- IP distortions
- H.264 compression
- MPEG-2 compression

All video sequences are 10 second long videos except one which is 8.68 seconds of video. Therefore we would have 250 frames in 25 frame per second videos and 500 frames in the 50 frame per second videos in the case of the video with a shorter length, since it is a 25 frame per second video we would have 217 frames in the sequence.

4.2 Calculating the data to be pooled

The first step in calculating the proposed method is finding the matching motion blocks and calculating the motion vectors from them. For doing this calculation we used a sub-block of size 16 × 16 pixels and gave the sub-block the possibility to move 7 pixels in each direction (in other words a search parameter of 7), shown in Figure 22. By finding the matching sub-blocks we would easily calculate the quality values of each channel and then calculate the values needed to be pooled.

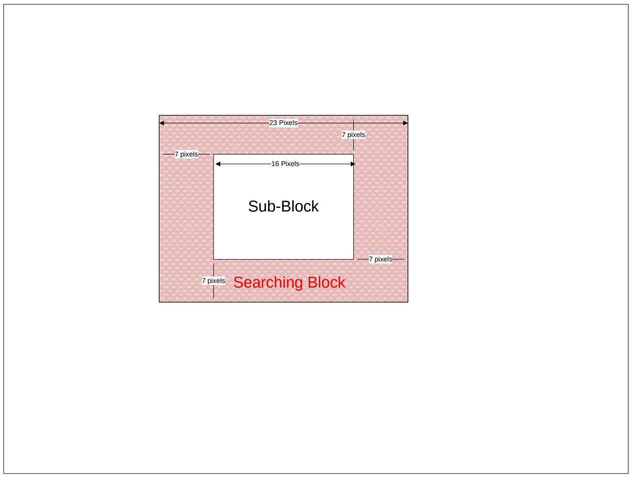


Figure 22. example of the micro-block and its search block in a frame.

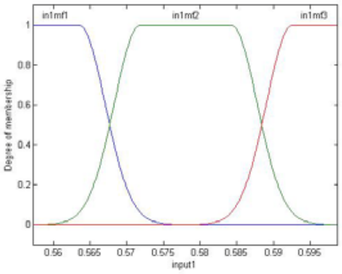
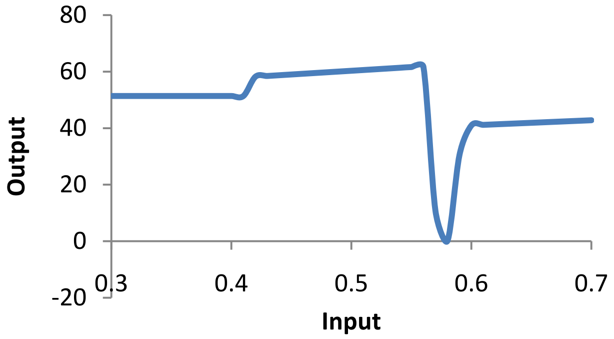
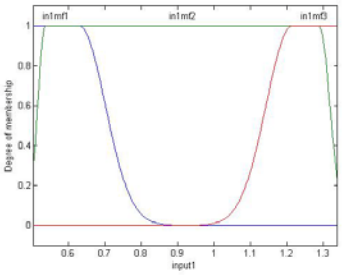
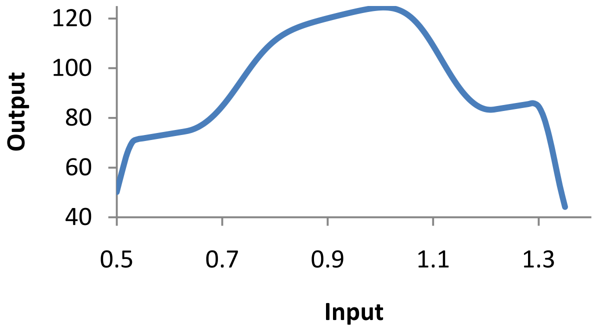
4.3 Pooling the data

For pooling the data we will first pool each distortion/compression separately and then try to pool all data’s at once. For pooling the quality of the videos of the database in each group we will have 5 different pooling functions each representing one motion activity density group. Each system is trained by around %60-70 of the data. The difference is because there are not the same number of videos in each distortion/compression group also the number of videos in each motion activity density group is different.

4.3.1 IP distortion

As mentioned before we would have 5 different pooling functions corresponding to each motion activity group. Membership function and the plot of input vs. output for different motion activity groups could be seen in Table 7.

Table 7. Membership function and the plot of input vs. output for the 5 different motion activity density groups for videos with IP distortion.

Motion activity group	Membership function	Output vs. Input
1		
2		

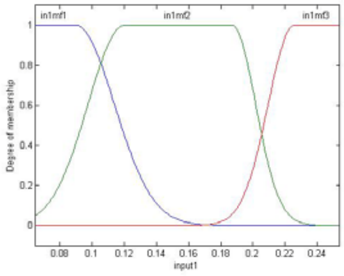
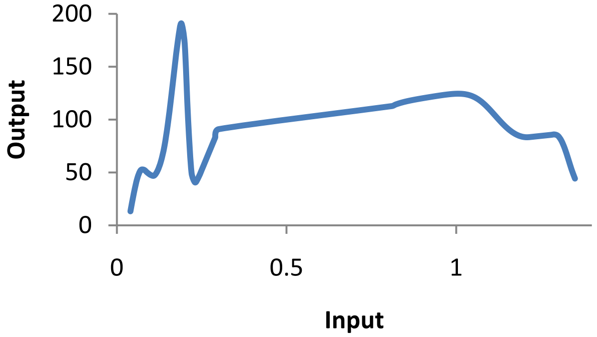
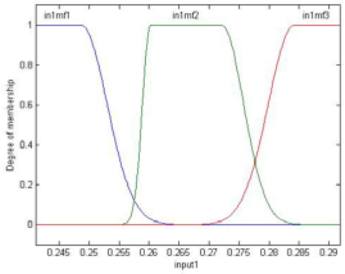
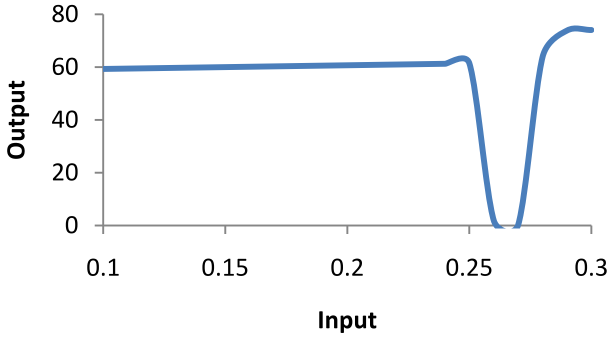
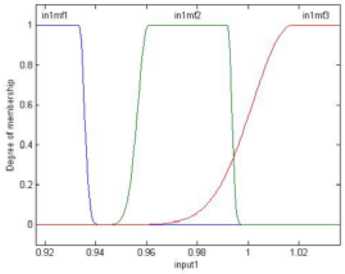
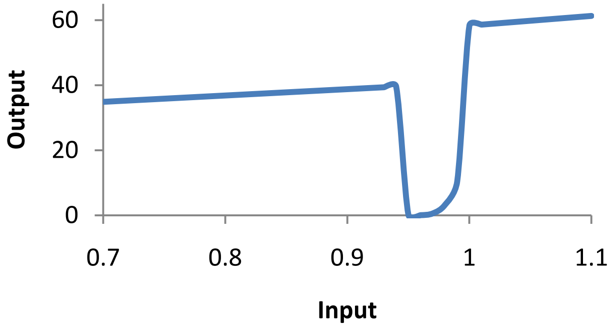
3		
4		
5		

Figure 23 shows the objective scores against the subjective scores for IP distorted videos.

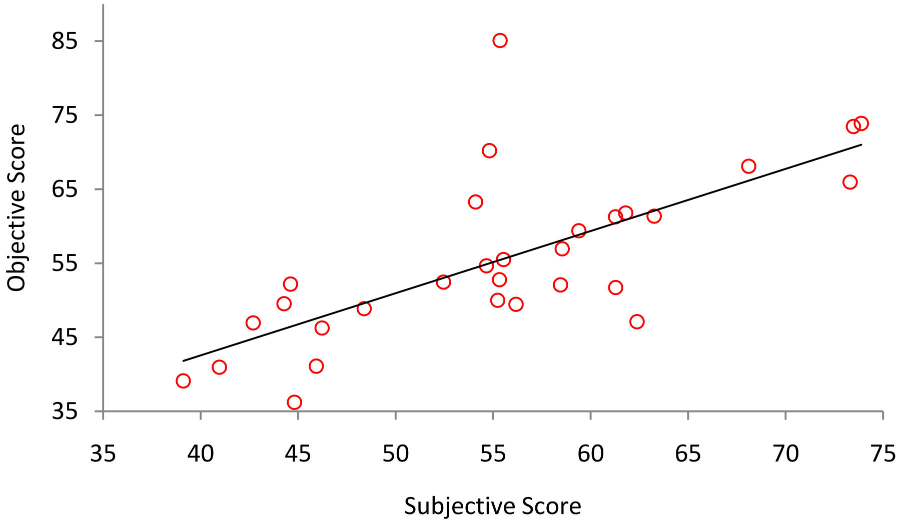


Figure 23. Subjective scores vs. Objective scores in IP distorted videos.

Table 8 shows the Pearson and Spearman correlation for 10 different FR state of the art VQAM's and also the results for the proposed RR metric. As it can be seen the proposed metric is ranked first in the Spearman correlation but ranked sixth in the Pearson correlation. The reason for that is that the correlation results are really close to each other for example in the case of Pearson correlation the difference between the proposed metric and the best metric is 5.42% and in the case of the Spearman correlation the proposed metric is giving just a 0.01% better result.

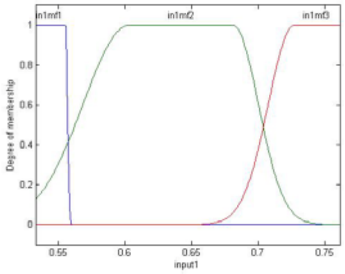
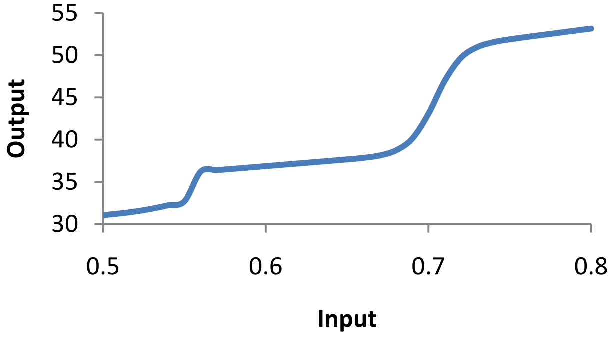
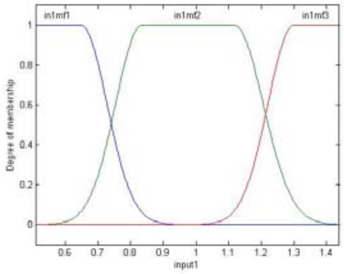
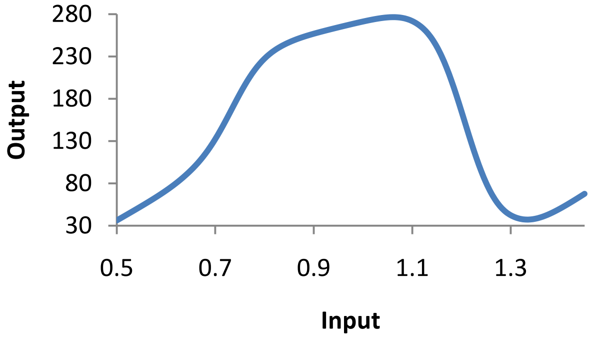
Table 8. Spearman and Pearson correlation results for 10 different state of the art VQAM and also the proposed metric in the IP distorted videos.

Prediction Model	Spearman Correlation	Pearson Correlation
PSNR	0.3206	0.4108
SSIM	0.4550	0.5119
MS-SSIM	0.6534	0.7219
Speed SSIM	0.4727	0.5587
VSNR	0.6894	0.7341
VQM	0.6383	0.6480
V-VIF	0.4736	0.5102
Spatial MOVIE	0.7046	0.7378
Temporal MOVIE	0.7192	0.7383
MOVIE	0.7157	0.7622
Proposed Method	0.7202	0.7080

4.3.2 H.264 compression

As like the IP distorted videos we will have 5 different pooling functions each corresponding to a motion activity density group. The Membership function and the plot of input vs. output for different motion activity groups could be seen in Table 9.

Table 9. Membership function and the plot of input vs. output for the 5 different motion activity density groups for H.264 compressed videos.

<table>
<tr><th>Motion activity group</th><th>Membership function</th><th>Output vs. Input</th></tr>
<tr><td>1</td><td>
</td><td>
</td></tr>
<tr><td>2</td><td>
</td><td>
</td></tr>
</table>

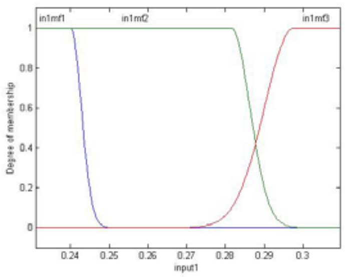
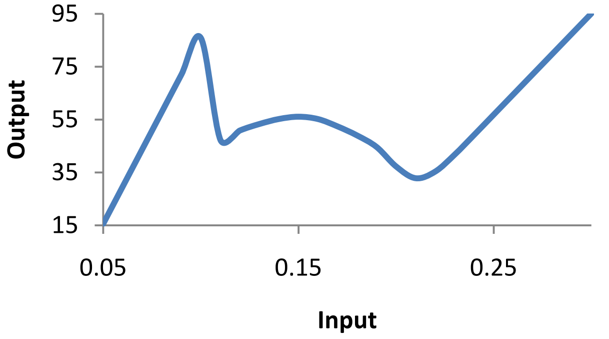
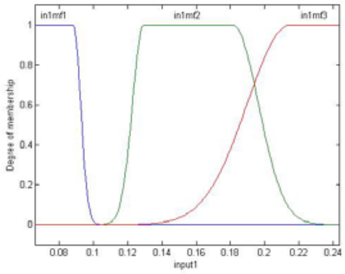
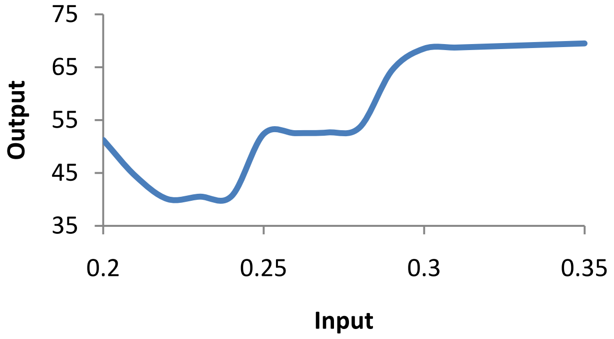
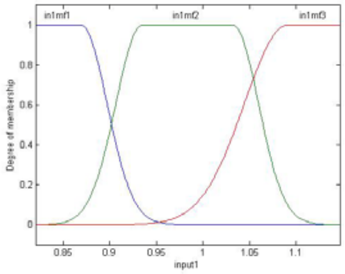
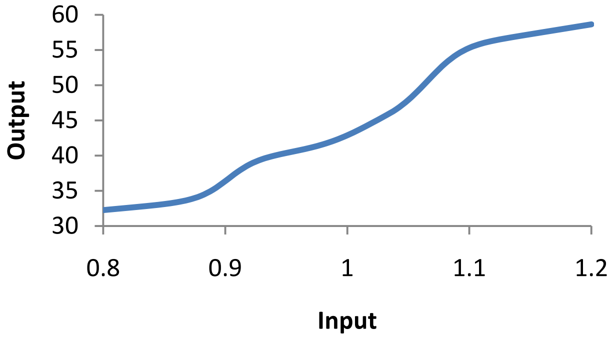
3		
4		
5		

Figure 24 shows the objective scores against the subjective scores for H.264 compressed videos.

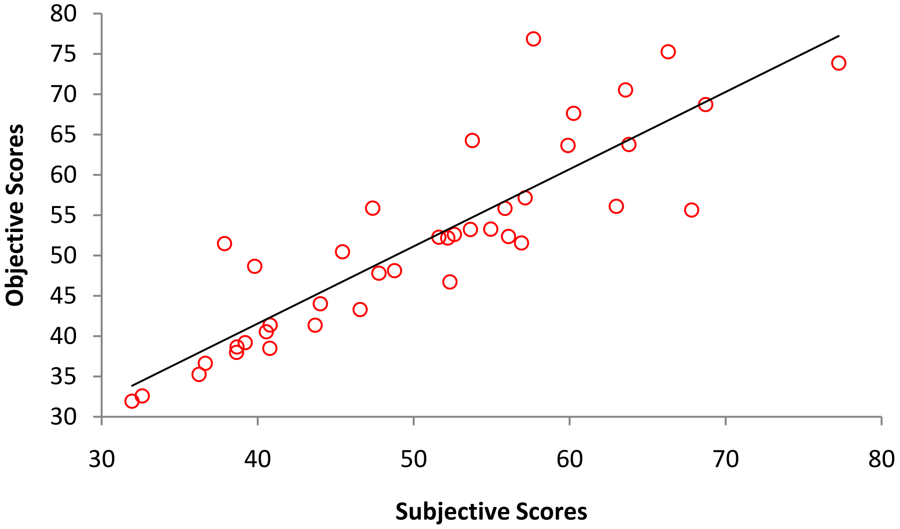


Figure 24. Subjective scores vs. Objective scores inH.264 compressed videos.

Table 10 shows the Pearson and Spearman correlation for 10 different FR state of the art VQAM's and also the results for the proposed metric. As it can be seen the proposed metric is ranked first in both the Spearman and Pearson correlation. In the case of Spearman correlation we are getting a result which is 13.35% better than the second best result and in the case of Spearman correlation we are getting results which are 8.58% better than the next result.

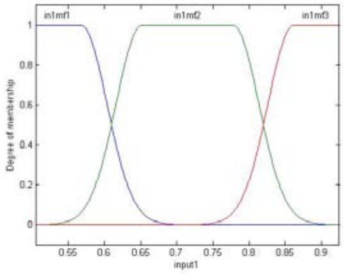
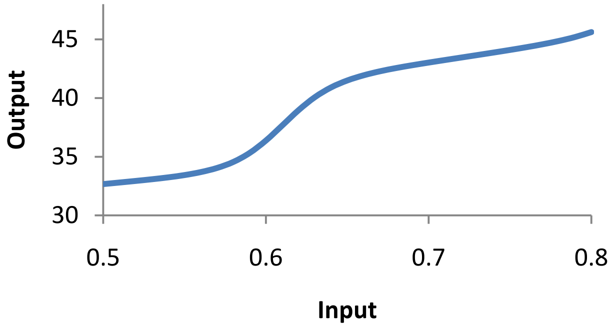
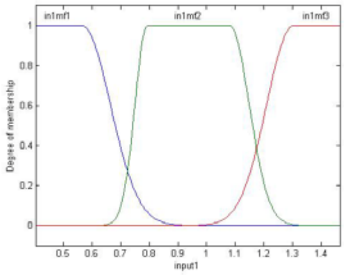
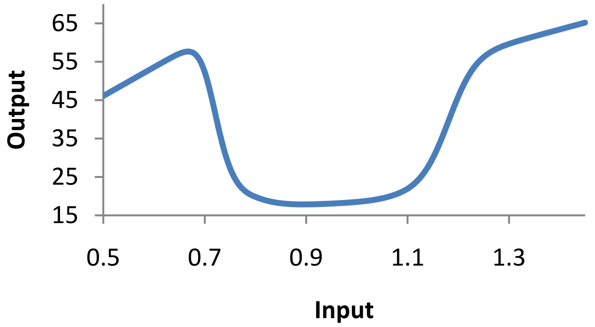
Table 10. Spearman and Pearson correlation results for 10 different state of the art VQAM and also the proposed metric in the H.264 compressed videos.

Prediction Model	Spearman Correlation	Pearson Correlation
PSNR	0.4296	0.4385
SSIM	0.6514	0.6656
MS-SSIM	0.7051	0.6919
Speed SSIM	0.7086	0.7206
VSNR	0.6460	0.6216
VQM	0.6520	0.6459
V-VIF	0.6807	0.6911
Spatial MOVIE	0.7066	0.7252
Temporal MOVIE	0.7797	0.7920
MOVIE	0.7664	0.7902
Proposed Method	0.9132	0.8778

4.3.3 MPEG-2 compression

As like the previous two approaches we will have 5 different pooling functions each corresponding to a motion activity density group. The Membership function and the plot of input vs. output for different motion activity groups could be seen in Table 11.

Table 11. Membership function and the plot of input vs. output for the 5 different motion activity density groups for MPEG-2 compressed videos.

Motion activity group	Membership function	Output vs. Input
1		
2		

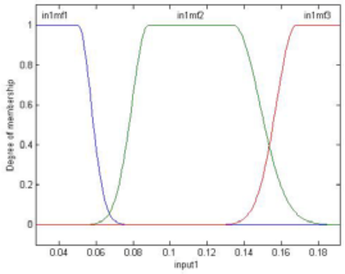
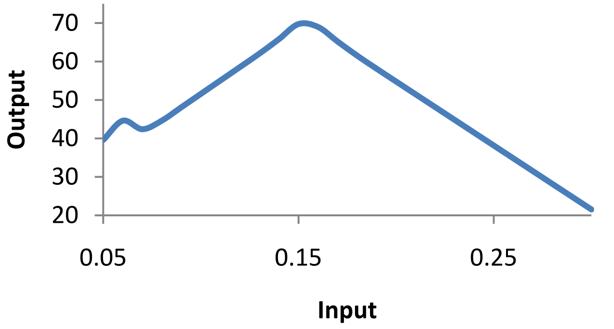
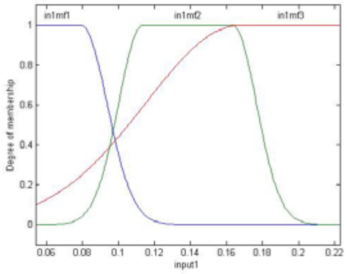
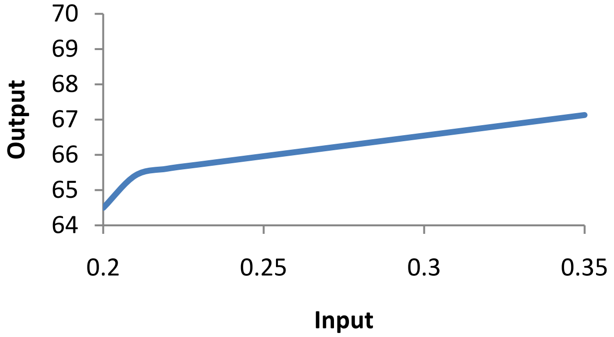
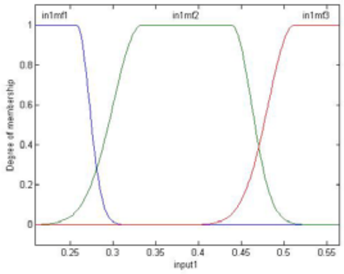
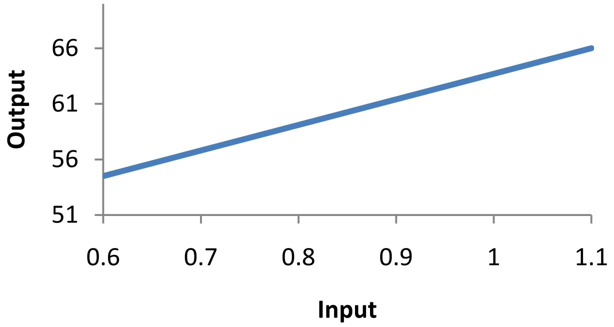
3		
4		
5		

Figure 25 shows the objective scores against the subjective scores for MPEG-2 compressed videos.

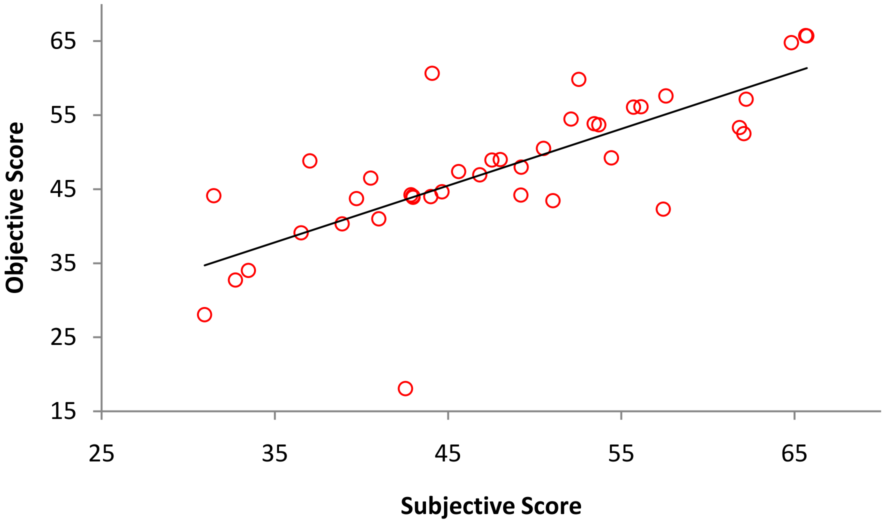


Figure 25. Subjective scores vs. Objective scores in MPEG-2 compressed videos.

Table 12 shows the Pearson and Spearman correlation for 10 different FR state of the art VQAM's and also the results for the proposed metric. As it can be seen the proposed metric is ranked third in the Spearman correlation and second in the case of Pearson correlation. In the case of the Spearman correlation our metric is giving results which are 5.18% less than the best metric in correlation and in the case of Pearson correlation we are behind the best metric by 2.64%.

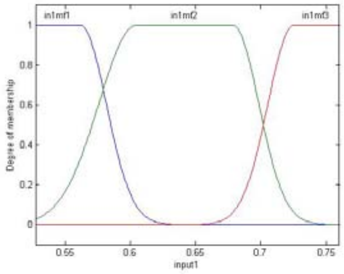
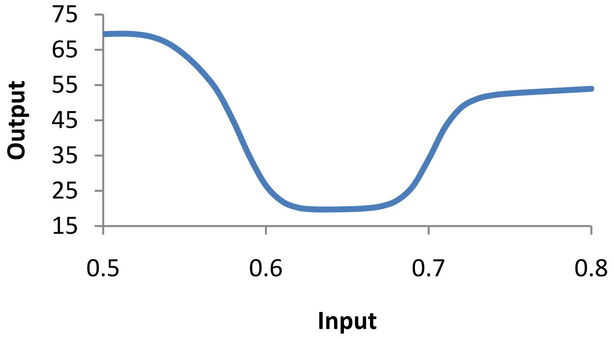
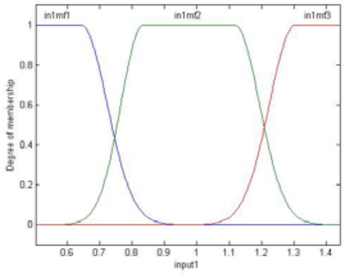
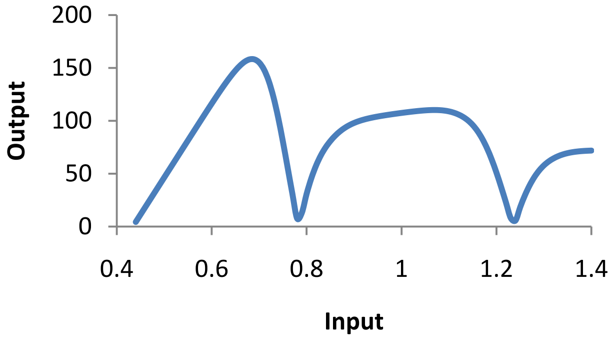
Table 12. Spearman and Pearson correlation results for 10 different state of the art VQAM and also the proposed metric in the MPEG-2 compressed videos.

Prediction Model	Spearman Correlation	Pearson Correlation
PSNR	0.3588	0.3856
SSIM	0.5545	0.5491
MS-SSIM	0.6617	0.6604
Speed SSIM	0.6185	0.6270
VSNR	0.5915	0.5980
VQM	0.7810	0.7860
V-VIF	0.6116	0.6145
Spatial MOVIE	0.6911	0.6587
Temporal MOVIE	0.8170	0.8252
MOVIE	0.7733	0.7595
Proposed Method	0.76515	0.7988

4.3.4 Wireless distortion

Although the metric proposed is based on block matching and so was not designed to deal with wireless distortions but we applied the metric on the wireless distorted videos as well to see how it would react to these types of distortions. Table 13 shows the membership functions and also the output vs. input plot of the pooling system.

Table 13. Membership function and the plot of input vs. output for the 5 different motion activity density groups for wireless distorted videos.

Motion activity group	Membership function	Output vs. Input
1		
2		

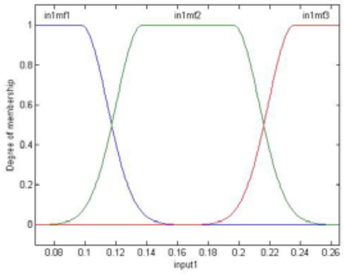
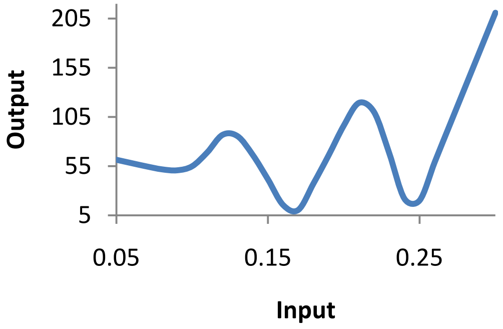
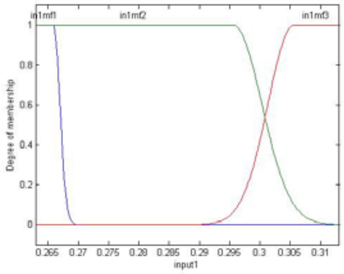
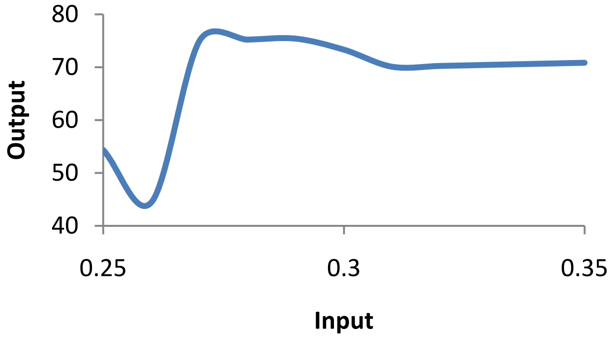
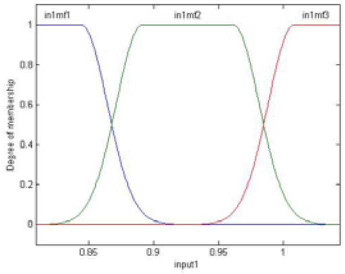
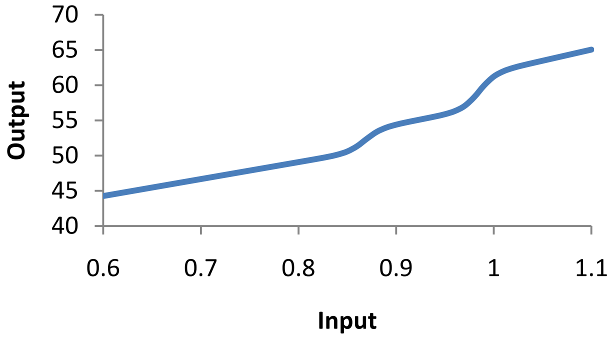
3		
4		
5		

Figure 26 shows the objective scores against the subjective scores for wireless distorted videos.

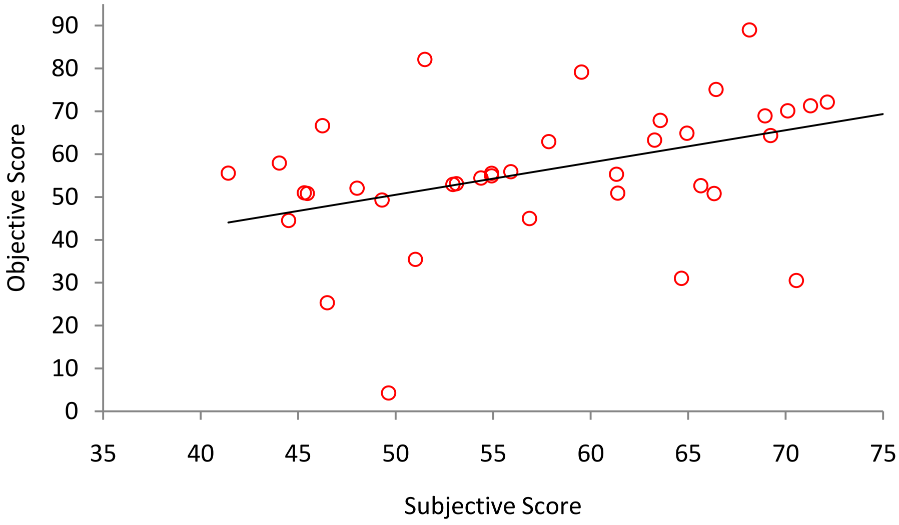


Figure 26. Subjective scores vs. Objective scores in the Wireless distortion videos.

Table 14 shows the Pearson and Spearman correlation for 10 different FR state of the art VQAM's and also the results for the proposed metric. As it can be seen the proposed metric does not give a high correlation when subjective scores are compared against the objective scores.

Table 14. Spearman and Pearson correlation results for 10 different state of the art VQAM and also the proposed metric in the wireless distorted videos.

Prediction Model	Spearman Correlation	Pearson Correlation
PSNR	0.4334	0.4675
SSIM	0.5233	0.5401
MS-SSIM	0.7285	0.7107
Speed SSIM	0.5630	0.5867
VSNR	0.7019	0.6992
VQM	0.7214	0.7325
V-VIF	0.5507	0.5488
Spatial MOVIE	0.7927	0.7883
Temporal MOVIE	0.8114	0.8371
MOVIE	0.8109	0.8386
Proposed Method	0.5394	0.4684

4.3.5 All the videos in the database

In the case when we try to find results for all different videos we will not get really good results as shown in Figure 27 which shows the objective scores against the subjective scores for all the videos. This is mainly because we are trying to train our system without taking the fact that they are being distorted or compressed in different ways which that will affect the whole system significantly. For example we are trying to evaluate the quality of a video which is compressed by an H.264 format with a pooling function which is trained based on data's which have wireless distortion.

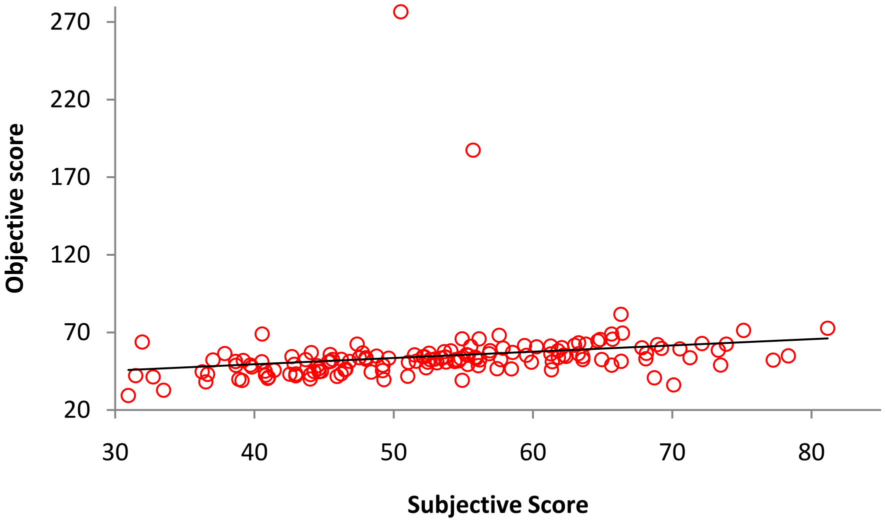
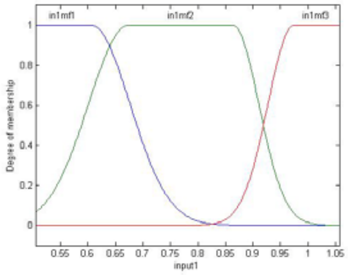
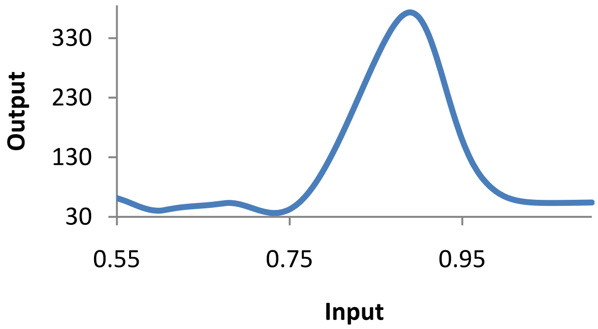
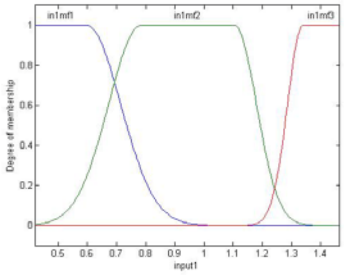
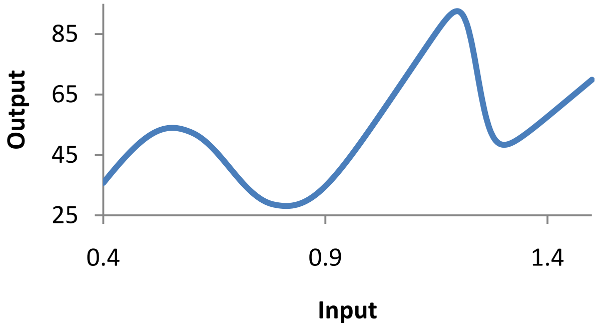
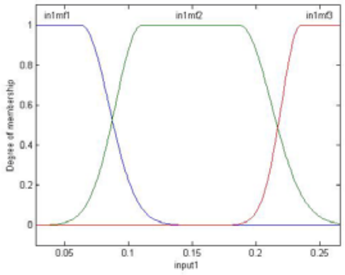
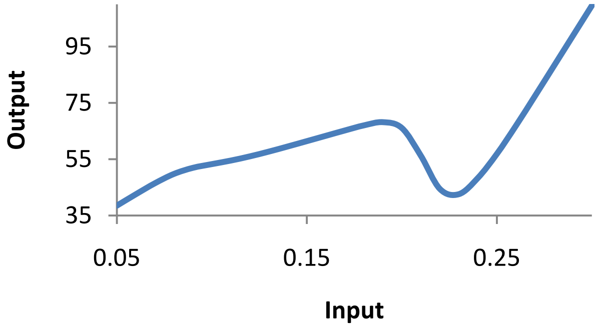


Figure 27. Subjective scores vs. Objective scores in all the videos in the database.

Table 15 shows the membership functions along with the plot of output vs. input data.

Table 15. Membership function and the plot of input vs. output for the 5 different motion activity density groups for all the videos.

Motion activity group	Membership function	Output vs. Input
1		
2		
3		

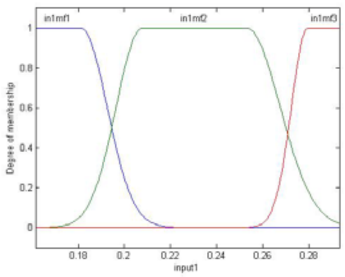
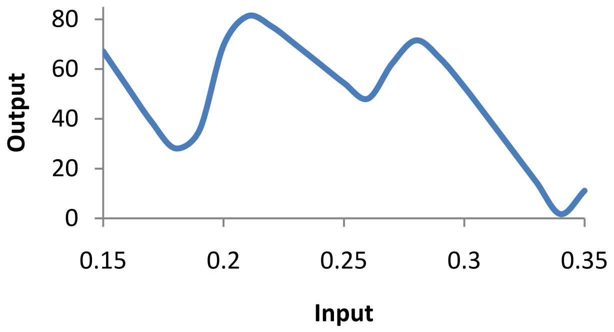
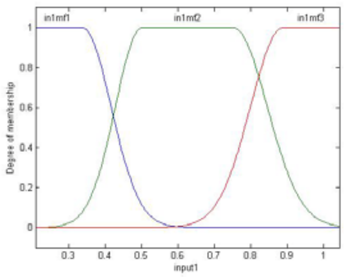
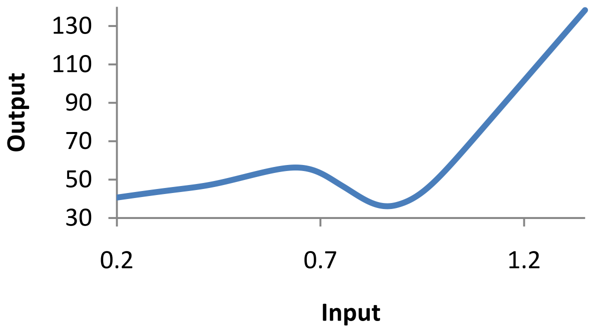
4		
5		

Table 16 shows the Spearman and Pearson correlation of the proposed metric along with 10 other state of the art FR VQAM's. As mentioned previously we do not get a high correlation among all the videos but this is because we are mixing and comparing videos which are from different types of distortions/compressions and our method was not designed for some of them.

Table 16. Spearman and Pearson correlation results for 10 different state of the art VQAM and also the proposed metric in all the videos.

Prediction Model	Spearman Correlation	Pearson Correlation
PSNR	0.3684	0.4035
SSIM	0.5257	0.5444
MS-SSIM	0.7361	0.7441
Speed SSIM	0.5849	0.5962
VSNR	0.6755	0.6896
VQM	0.7026	0.7236
V-VIF	0.5710	0.5756

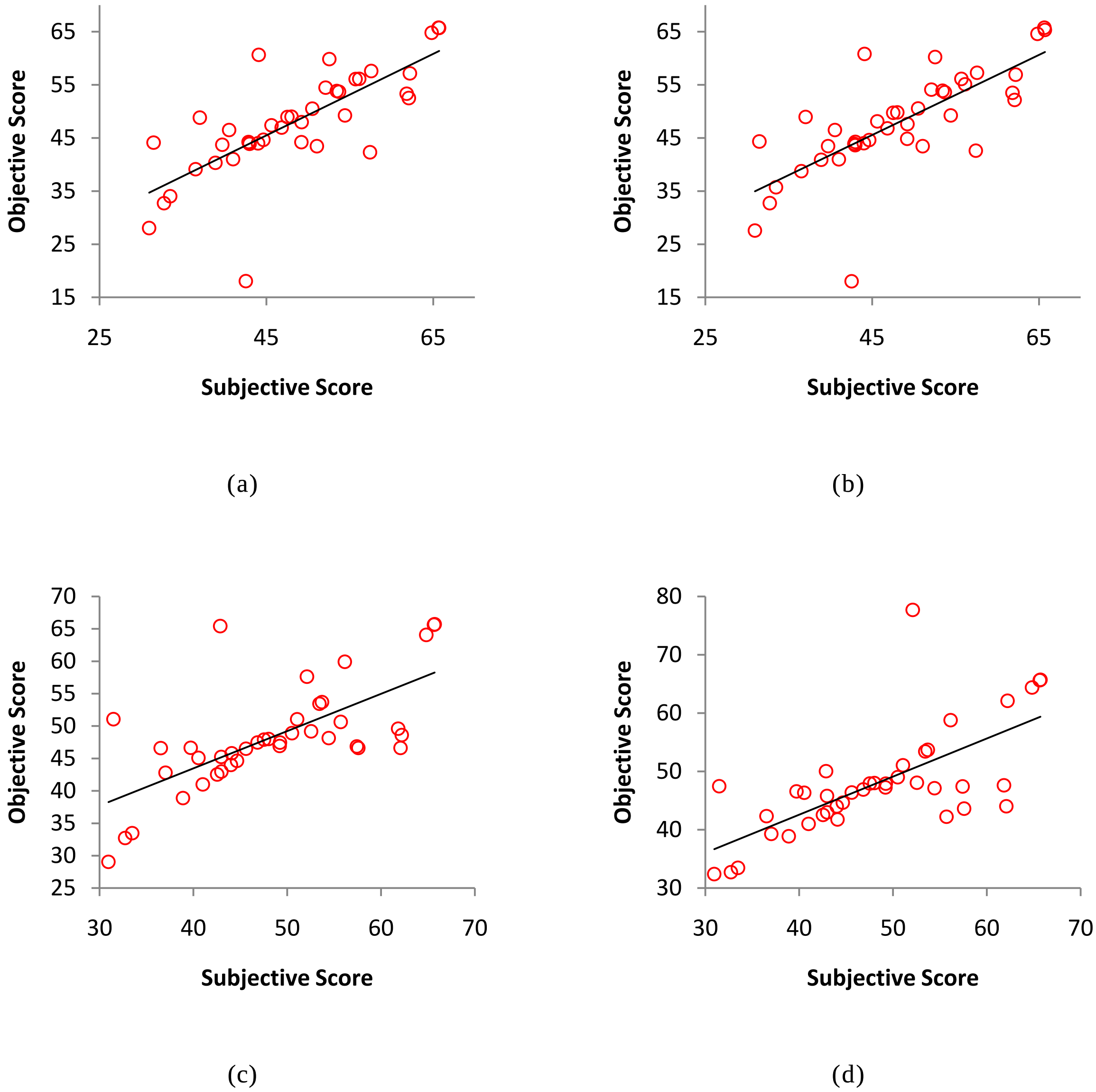
Spatial MOVIE	0.7270	0.7451
Temporal MOVIE	0.8055	0.8217
MOVIE	0.7890	0.8116
Proposed Method	0.5665	0.194

4.4 Evaluating other factors

Apart from pooling the data according to the proposed method we also pooled the data with the three following changes in the input data to check the effect different factors might have in our overall quality score:

1. In order to evaluate the role of the attention map on the overall score we neglected using the weighting procedure on the luminance channel using the Visual Attention Map (VAM) scores.
2. In order to evaluate the role of the matching motion vector value in our method we neglected using this factor in the original scheme and evaluated the results.
3. In the last test we subtracted the matching vector factor from one or in other words we used the number of nonmatching MV's to see how our method would react to this change.

Figure 28, Figure 29, Figure 30, Figure 31 and Figure 32 show the related plots of the subjective score against the objective score for all the 4 different approaches and all the distorted/compressed video.



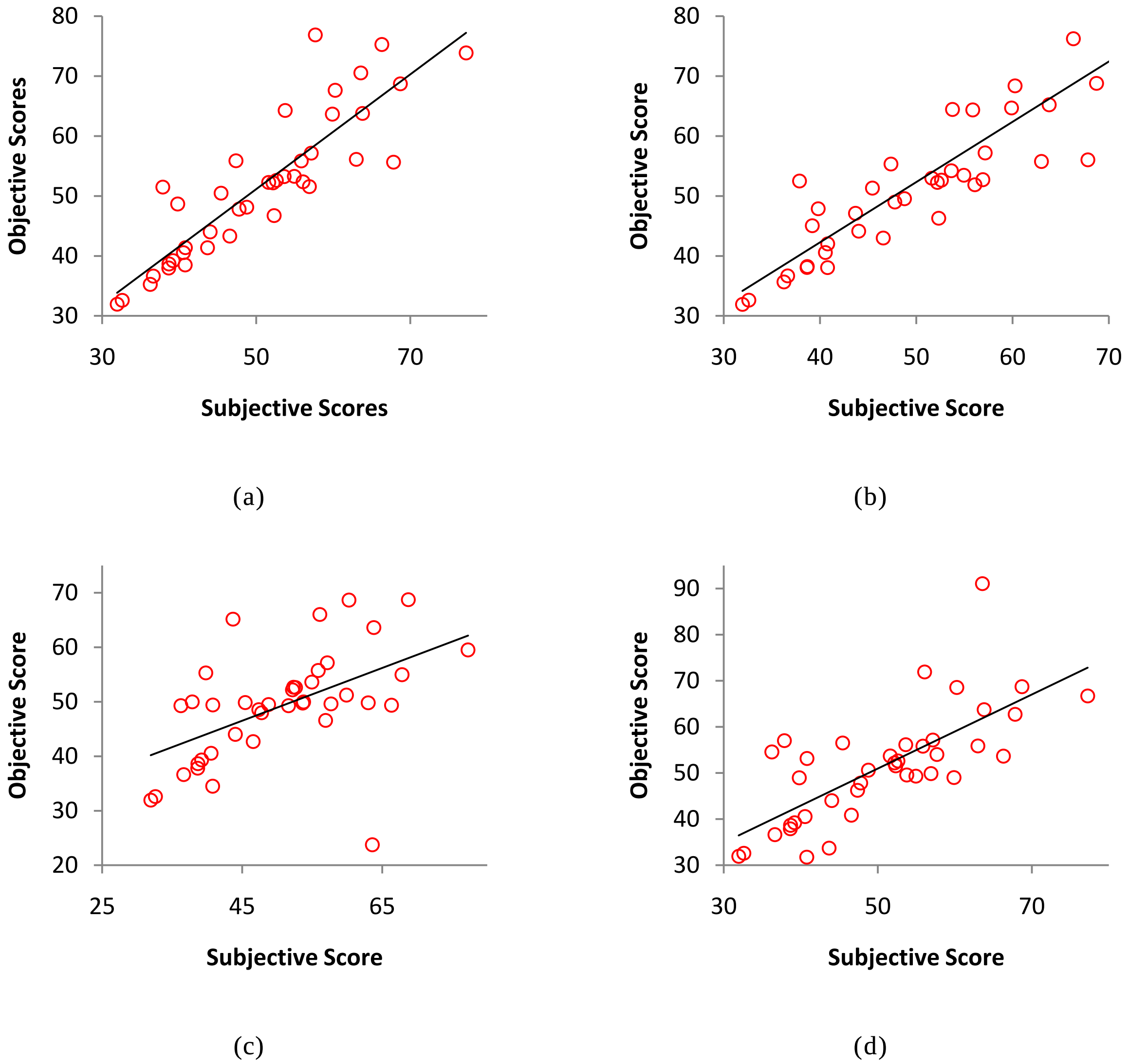
(a)

(b)

(c)

(d)

Figure 28. Subjective score vs. Objective score for different approaches in the MPEG-2 compressed videos (a) proposed approach (b) without using the attention map (c) subtracting the matching MV value from one (d) not using the matching MV value.



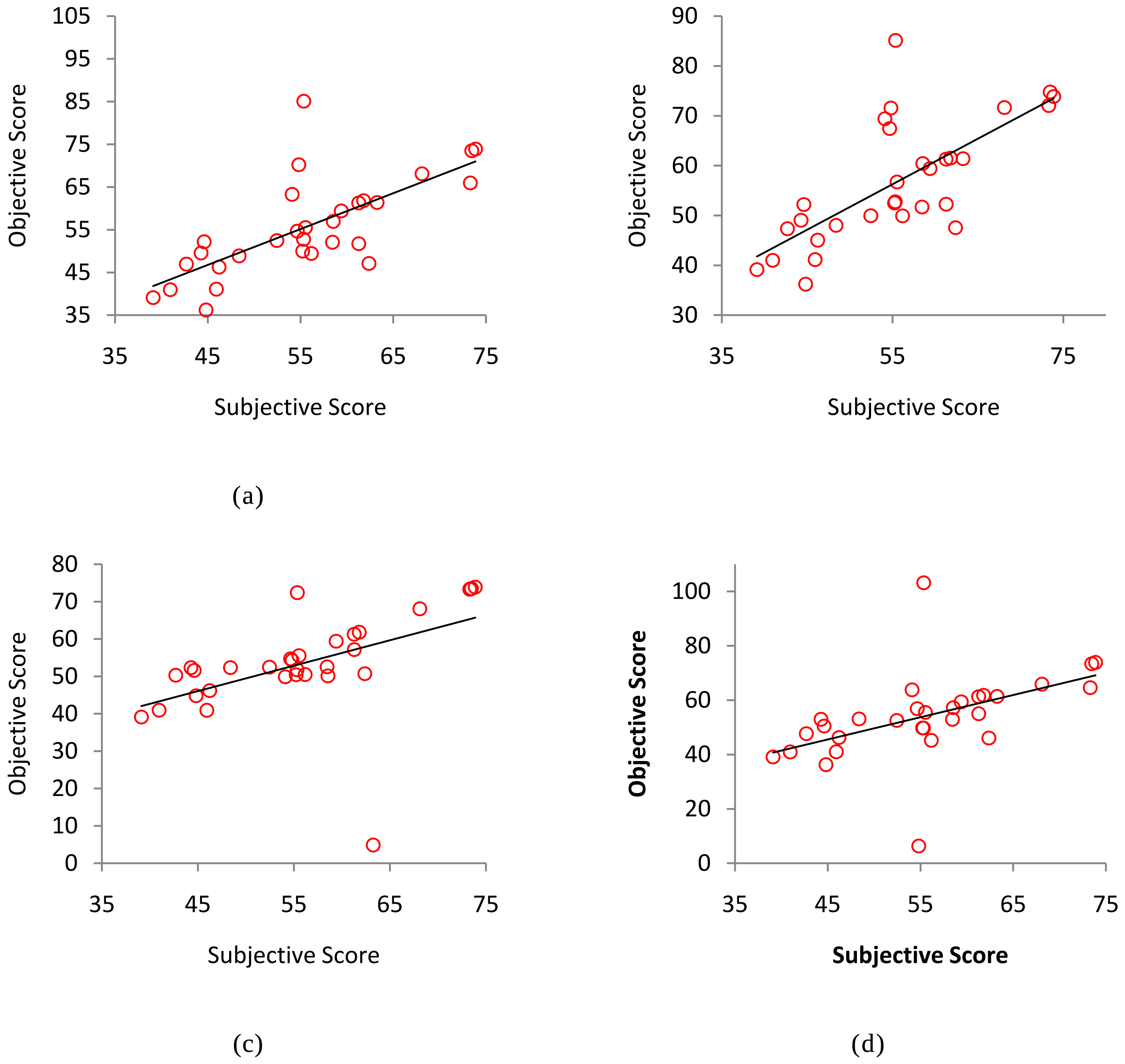
(a)

(b)

(c)

(d)

Figure 29. Subjective score vs. Objective score for different approaches in the H.264 compressed videos (a) proposed approach (b) without using the attention map (c) subtracting the matching MV value from one (d) not using the matching MV value.

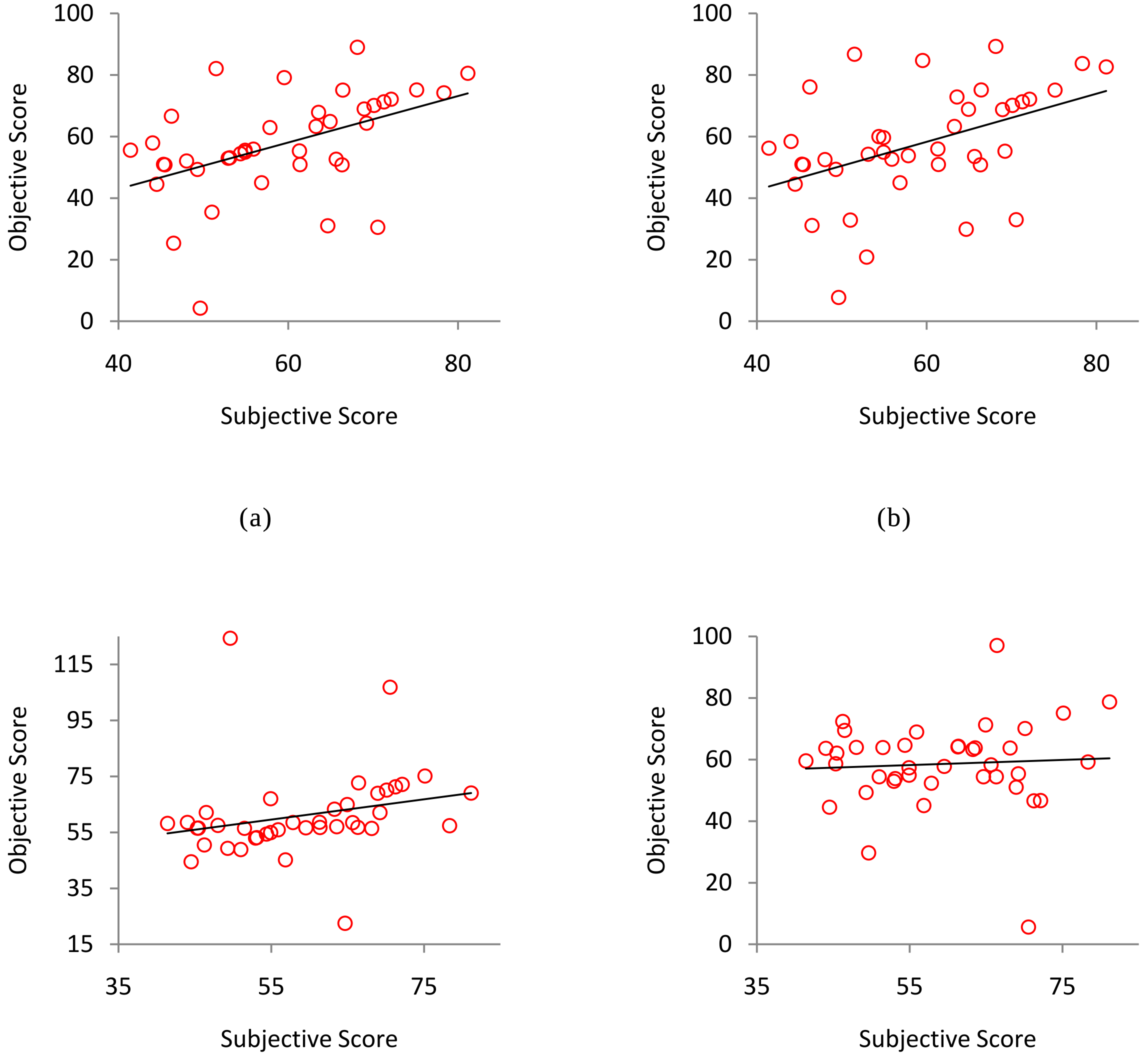


(a)

(c)

(d)

Figure 30. Subjective score vs. Objective score for different approaches in the IP distorted videos (a) proposed approach (b) without using the attention map (c) subtracting the matching MV value from one (d) not using the matching MV value.



(a)

(b)

(c)

Figure 31. Subjective score vs. Objective score for different approaches in the wireless distorted videos (a) proposed approach (b) without using the attention map (c) subtracting the matching MV value from one (d) not using the matching MV value.

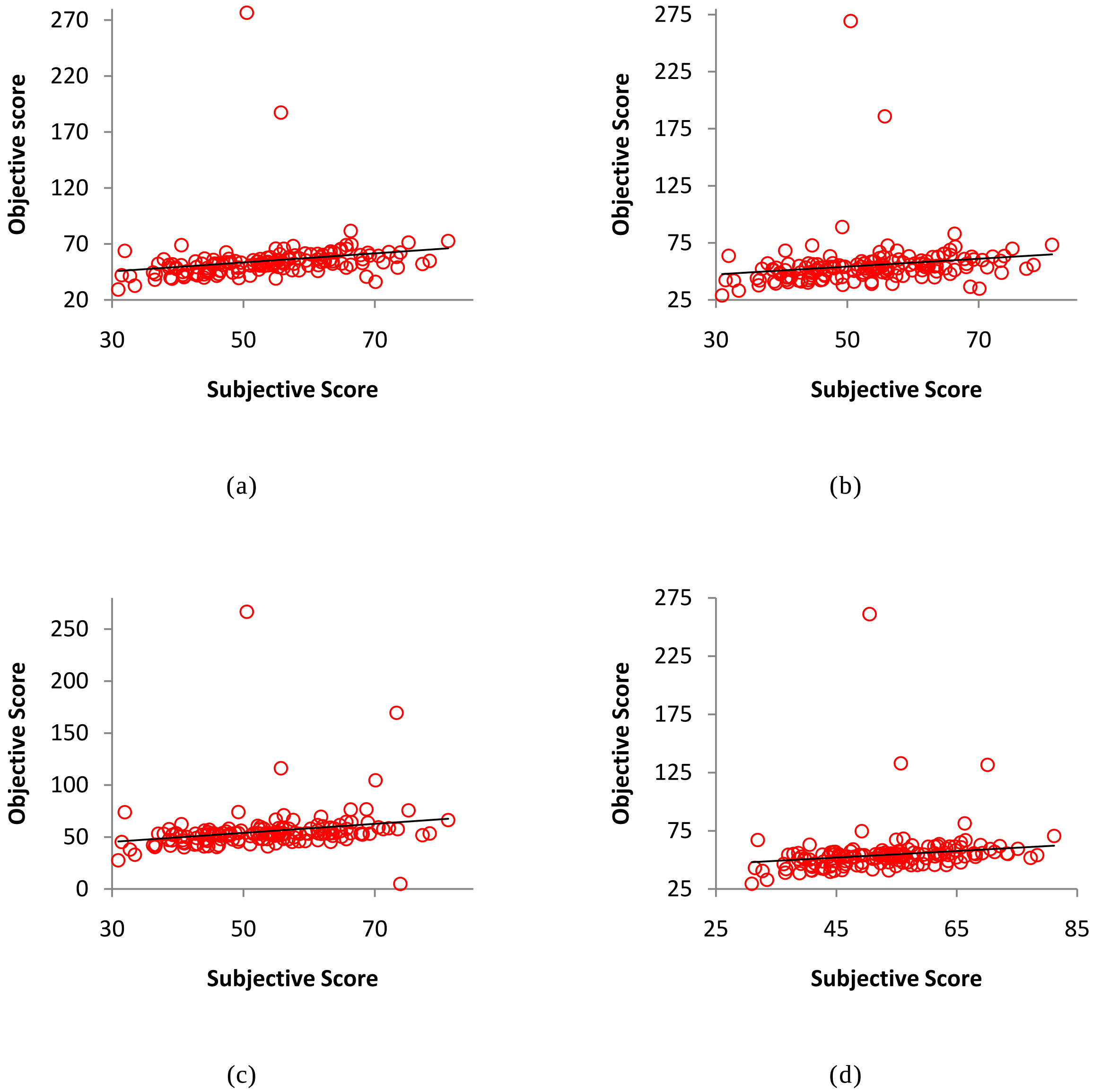


Figure 32. Subjective score vs. Objective score for different approaches in all the videos (a) proposed approach (b) without using the attention map (c) subtracting the matching MV value from one (d) not using the matching MV value.

To compare the different approaches we will calculate the Spearman correlation and also the Pearson correlation for different approaches which is shown in Table 17 and Table 18. As it can be seen any change in the matching MV score will affect the overall score and therefore reduce the precision of the proposed method. Also using the VAM will increase the overall score by a couple of percent. This along with the fact that calculating the VAM is a fast and easy process encourages us to keep this factor in our method. Overall with respect to all proposed approaches the main proposed approach will give us the best results.

Table 17. Spearman correlation of different approaches.

Approach taken	Wireless	IP	H.264	MPEG-2	All Data
Proposed method	0.5394	0.7202	0.9132	0.76515	0.5665
Without using an VAM weighting system	0.4805	0.7290	0.9094	0.7965	0.4579
Subtracting the matching MV score form one	0.5434	0.6746	0.6101	0.7138	0.4766
Without using the matching MV score	0.1301	0.7064	0.7341	0.7024	0.4604

Table 18. Pearson correlation of different approaches.

Approach taken	Wireless	IP	H.264	MPEG-2	All Data
Proposed method	0.4684	0.7080	0.8778	0.7988	0.194
Without using an VAM weighting system	0.4445	0.7119	0.8537	0.7462	0.1667
Subtracting the matching MV score form one	0.2386	0.4910	0.5308	0.681	0.2104
Without using the matching MV score	0.062	0.4936	0.7257	0.6853	0.143

Another important factor which could be easily seen in Table 17 and Table 18 and also in Figure 32 is the low correlation score when we want to evaluate the quality of all videos at once. This is not only because of the fact that we are comparing different videos from totally different distortion/compression types but also because a limited number of videos seem to be giving really bad scores. If we remove these videos and calculate the correlation again we will see a good increase in the overall correlation value which is shown in Table 19.

Table 19. correlation for all the videos.

Prediction Model	Pearson Correlation	Spearman Correlation
Before removing the wrong values	0.194	0.5665
After removing the wrong values	0.5563	0.5767

5 Conclusion and future works

5.1 Conclusion

In this work we have introduced a new Reduced Reference Video Quality Metric which is fundamentally based on block matching. Because of the spatial-temporal approach we have taken the method is named as STAQ (Spatial-Temporal Assessment of Quality). For each sub-block in a frame we will find the matching sub-block in the next frame of the video and take the next frame's matching sub-block as the reference for the sub-block in the current frame. We will then use a simple Image Quality Metric to evaluate the quality in different channels and later on give a weight to each sub-block in the luminance channel according to the Visual Attention Map calculated for that specific frame. We will also compare the Motion Vectors in the reference video with the MV's in the test video and take the calculated value as a factor in our pooling system. The videos are also grouped based on the Motion Activity Density they have, into five different groups and each group is trained with the data we have from the same group. The pooling system is a neuro-fuzzy system and is implemented using the anfis toolbox of Matlab.

As it was demonstrated in Chapter 4 the best results were in the case of H.264 compressed videos with correlation values higher than other state of the art FR metrics. Also in the MPEG-2 videos we are getting results which are good and are running behind a metric which was proposed a couple of months ago [**28**]. In the case of IP distortion we are not ranked high in the case of Pearson correlation but this is mainly because the resulted values among different metrics are really close. In fact there is only a difference of %6 between our metric and the best metric while in the case of H.264 compression we have nearly % 15 improvement between the correlation value of our metric and the next metric. We do not get good results in the case of wireless distortion and also when all the videos are compared with each other. Keeping in mind the progress made in other fields and also the fact that wireless distortion should be treated totally differently we could claim that the proposed metric is among the best metrics introduced so far.

A reason for getting good results in the case of H.264 and MPEG-2 compressed videos might be the fact that we also have a block matching approach in our method which is also used when the videos are compressed. On the other hand we are facing distortions such as bleeding, ringing and blurring in the case of wireless distortion which is something that it is not easy to calculate and model using the block matching approach we are taking.

5.2 Future work

Regarding the future works that can be done to improve the proposed metric and have a metric which is more depended on the HVS and perceptual observation and to have a greater focus on QOE rather than QOS there are a few points that could be useful.

One of the most important points that should be taken into account is that now days you cannot neglect the presence of audio in most videos. Apart from some silent movies all videos are accompanied with some kind of audio. This audio will affect the observer's attention as well so normal saliency region detection methods and their corresponding attention maps would not give us correct result when applied on videos which have an audio. This will also affect the VQM's which use some sort of attention maps (like our proposed method).

Since the proposed approach is depended on the matching sub-blocks and the MV's, we could use the motion activity density group when calculating the MV's and also the threshold for the accepted sub-blocks. Figure 33 shows a schematic of how the method would look like, the red parts are the parts added compared to the proposed approach. To give a better explanation, instead of giving a fixed value for the search parameter (in our case 7) we would select the value of the search parameter depending on the motion activity density group the video belongs to. For videos with low motion activity density this value would be low and for videos with high motion activity density this value will be higher. Also in the case of selecting the threshold we would assign a higher threshold and so a wider range for the accepted MV's in videos with high motion activity density and a lower threshold and so a lower range for the MV's of videos whom belong to a lower motion activity density group. This way we will not only increase the precision of our results but we will reduce the calculation time as well.

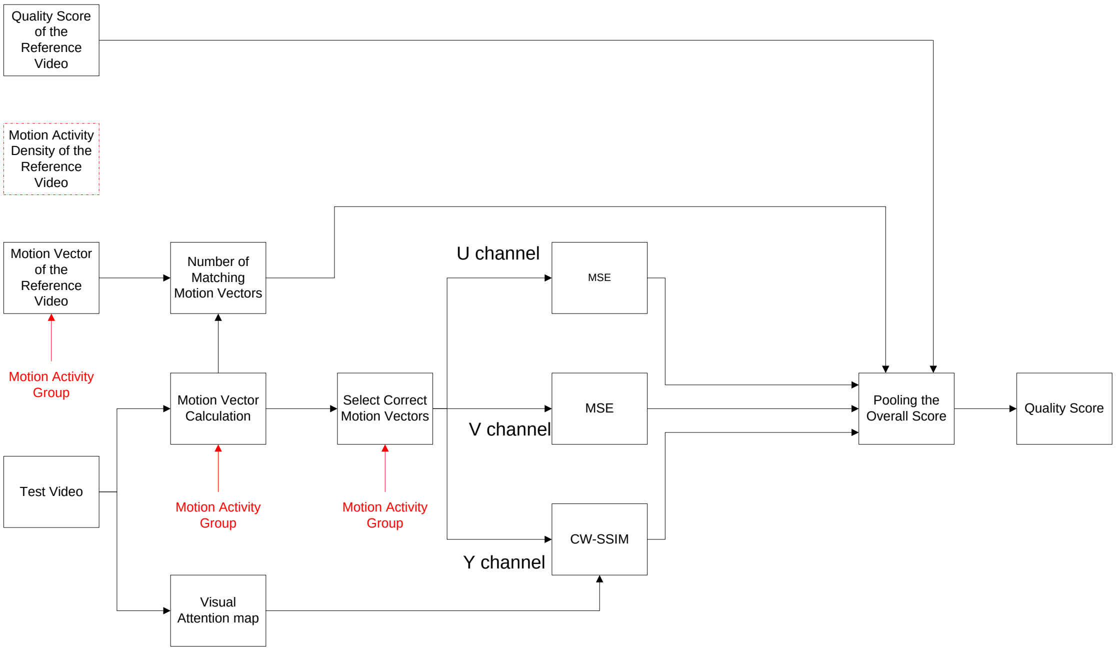

Quality Score of the Reference Video
Motion Activity Density of the Reference Video
Motion Vector of the Reference Video
Number of Matching Motion Vectors
U channel
MSE
Motion Activity Group
Motion Vector Calculation
Select Correct Motion Vectors
V channel
MSE
Pooling the Overall Score
Quality Score
Test Video
Motion Activity Group
Motion Activity Group
CW-SSIM
Y channel
Visual Attention map

Figure 33. schematic of the changes made.

Also a good approach could be giving different weights to different factors. Obviously the quality in the luminance channel has a higher importance than the quality in the chrominance channels. And the percentage of the matching MV's does not have the same importance as the other factors calculated. The important factor in giving a weight to different factors is that these weights should represent the real importance each have in the HVS.

Another important factor to work on would be the influence of the burst-of-error on the overall quality of the video. There is always a big question regarding how we should treat a video which has the same amount of quality in every frame compared to a video which the quality of the frames changes in a rather fast manner.

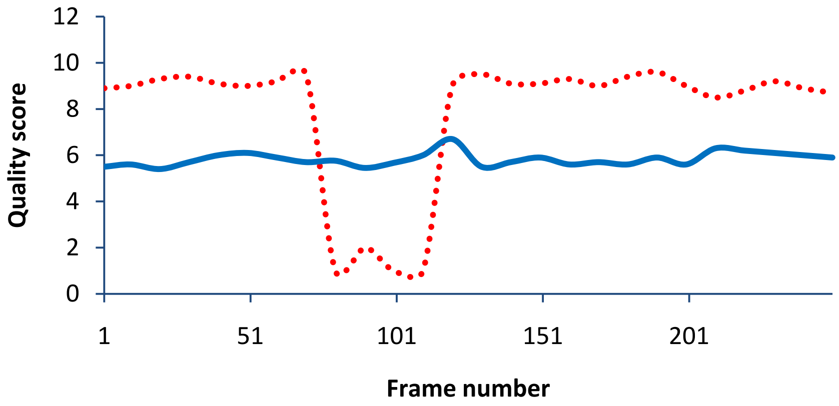


Figure 34. example of two video sequences, there is a burst-of-error in the case of the video shown in a dashed pattern.

An interesting factor in most VQAM's is that nearly all metrics give a single value as the output quality of a video. In my opinion giving just a single value could not give a clear description of the video quality. Giving results in a limited number of fields could be a better idea. Also instead of assigning a value to a single or a limited number of fields it could be better if we have a limited number of groups which will describe the quality of video and then determine which group the video belongs to. For instance could we really say that an observer could distinguish that a video with a quality value of 0.73 has a different quality compared to a video with a value of 0.77 in the case when we are using a metric which will give us values between zero and one? In the next step we could assign a membership value determining that in what rate each video belongs to a particular group.

Also the use of a Visual Attention Map which is designed for videos could give a better result than using VAM's that are designed for images. Since the observer would not have an equal time spent on a frame compared to the time he/she might spend on an image a VAM which is designed for videos would give us a better result.

For the case of wireless distortion (or even in other distortion types) it might be a good idea if we can detect the type of distortion we are facing and then have a slight change, or even take a different approach if we see that the main metric does not work correctly and with the same precision we have in other distortion/compression types.

In our opinion and with respect to all the video quality databases available none of them have been kept up to date with the progress made in the field of VQA. For example although the effect of burst-of-error has been mentioned in some publications [**17**] but there has not been any work done on this particular aspect. This might be because there is no database with such particular aspect among its videos. Furthermore there should be a database which has videos from a wide range of motion activity density. A database could also have the saliency regions detected by observes so that a full functional state of the art VQM could be proposed and tested on.

Bibliography

[1] A. M. Eskicioglu, "Quality measurement for monochrome compressed images in the past 25 years," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2000, pp. 1907-1910.

[2] Z. Wang, H. R. Sheikh and A. C. Bovik, *The handbook of video databases: design and application, Chapter 41: Objective video quality assessment*, B. F. a. O. Marques, Ed. CRC Press, 2003.

[3] P. V. Pahalawatta and A. M. Tourapis, "Motion estimated temporal consistency metrics for objective video quality assessment," in *International Workshop on Quality of Multimedia Experience, QoMEx*, 2009.

[4] S. Winkler and P. Mohandas, "The evolution of video quality measurement: from PSNR to hybrid metrics," *IEEE TRANS. BROADCASTING*, vol. 54, no. 3, pp. 1-9, Sep. 2008.

[5] L. Haglund. (2010, Jun.) The SVT high definition multi format test set. [Online]. ftp://vqeg.its.bldrdoc.gov/

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

[7] (2010, Apr.) Video Quality Experts Group (VQEG). [Online]. http://www.vqeg.org

[8] T. Brandão and P. Queluz, "No-reference perceptual quality metric for H.264/AVC encoded video," in *International Workshop on Video Processing and Processing and Quality Metrics for Consumer Electronics*, 2010.

[9] F. Yang, S. Wan, Y. Chang and H. R. Wu, "A novel objective No-Reference metric for digital video quality assessment," *IEEE Signal Processing Letters*, vol. 4, no. 11, pp. 685-688, Oct. 2005.

[10] Z. Wang, A. C. Bovik and B. L. Evans, "Blind measurment of blocking artifacts in images," in *IEEE*

International Conference on Image Processing (ICIP), Vancouver, 2000, pp. 981-984.

[11] H. R. wu and M. Yuen, "A generalized block-edge impairment metric for video coding," *IEEE signal processing letter*, vol. 4, no. 11, pp. 317-320, Nov. 1997.

[12] J. Caviedes and F. Oberti, "No-reference quality metric for degraded and enhanced video," in *SPIE*, 2003, pp. 621-632.

[13] P. Marziliano, F. Dufaux, S. Winkler and T Ebrahimi, "A no-reference perceptual blur metric," in *ICIP*, 2002, pp. 22-25.

[14] S. Winkler, *Digital video quality vision models and metrics*. John Wiley & Sons, 2005.

[15] B. Girod, *Digital images and human vision*. MA, USA: MIT Press, 1993.

[16] Q. huynh-Thu and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment," *Electronic Letters*, vol. 44, no. 13, Jun. 2008.

[17] Z. Wang, l. Lu and A. C. Bovik, "Video quality assessment based on structural distortion measurment," *Signal Processing: Special Issue on Objective Video Quality Metric*, vol. 19, no. 2, pp. 121-132, Feb. 2004.

[18] (2010, Apr.) Laboratory for Image and Video Engineering (LIVE) at The University of Texas at Austin. [Online]. http://live.ece.utexas.edu/

[19] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "Image quality assessment: from error visibility to stractural similarity," *IEEE Transaction on Image Processing*, vol. 13, no. 4, pp. 1-4, Apr. 2004.

[20] f. Zhang, J. Li, G. Chen and J. Man, "Assessment of color video quality with singular value decomposition of complex matrix," in *International Conference on Information Assurance and Security*, 2009.

[21] F. Zhang, "Quaternions and matrices of quaternions," *Linear algebra and its applications*, vol. 251, pp. 21-57, Jan. 1997.

[22] A. Shnayderman, A. Gusev and A. M. Eskicioglu, "An SVD-based gray-scale image quality measure for local and global assessment," *IEEE Transaction on Image Processing*, vol. 15, no. 2, pp. 422-429, Feb. 2006.

[23] N. Ohta and A. R. Robertson, *Colorimetry Fundamentals and Applications*. John Wiley & Sons, 2005.

[24] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. Prentice Hall, 2002.

[25] D. C. lin, P. M. Chau, "Objective human visual system based video quality assessment metric for low bit-rate video communication systems," in *IEEE Workshop on Multimedia Signal Processing*, 2006.

[26] E. P. Ong, X. Yang, E. Lin, Z. lu, S. Yao, x. Lin, S. Rahardja and B. C. Seng, "Perceptual quality and objective quality measurements of compressed videos," *Journal of Visual Communication and Image Representation*, vol. 17, no. 4, pp. 717-737, Aug. 2006.

[27] C. Kiemel and K. Diepold, "Improving the prediction accuracy of PSNR by simple temporal pooling," in *International Workshop on Video Processing and Quality Metrics (VPQM)*, 2010.

[28] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Transaction on Image Processing*, vol. 19, no. 2, pp. 335-350, Feb. 2010.

[29] A. Albonico, G. Valenzise, M. Naccari, M. Tagliasacchi and S. Tubaro, "A reduced-reference video structural similarity metric based on no-reference estimation of channel-induced distortion," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan, 2009.

[30] Y. Fu-zheng, W. Xin-dai, C. Yi-lin and W. Shuai, "A no-reference video quality assessment method based on digital watermark," in *International Symposium on Personal, Indoor and Mobile Radio Communication Proceedings.*, 2003.

[31] M. C. Q. Farias, M. Carli and S. K. Mitra, "Objective video quality metric based on data hiding," *IEEE transaction on Consumer Electronics*, vol. 51, no. 3, pp. 983-992, Aug. 2005.

[32] C. Opreal, I. Pirnog, C. Paleologu and M. Udrea, "Perceptual video quality assessment based on salient region detection," in *International Conference on Telecommunication*, 2009.

[33] M. Z. Aziz, B. Mertsching, "Fast and robust generation of feature maps for region-based visual attention," *IEEE transaction on Image Processing*, vol. 17, no. 5, pp. 633-644, Mar. 2008.

[34] A. Maalouf and M. C. Larabi, "A No-Reference color video quality metric based on a 3d multispectral wavelet transform".

[35] R. Achanta, S. Hemami, F. Estrada and S. Süsstrunk, "Frequency-tuned Salient Region Detection," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2009.

[36] A. Barajataya, "Block matching algorithms for motion estimation," ECE, Utah State University, Report for DIP course, 2004.

[37] Y. Nie, K. Ma, "Adaptive rood pattern search for fast block-matching motion estimation," *IEEE Transaction on Image Processing*, vol. 11, no. 12, pp. 1442-1449, Dec. 2002.

[38] M. P. Sampat, Z. Wang, A. c. Bovik and M. K. Markey, "complex wavelet structural similarity: a new image similarity index," *IEEE Transaction on Image Processing*, vol. 18, no. 11, pp. 2385-2401, Nov. 2009.

[39] K. Seshadrinathan, R. Soundararajan, A. C. Bovik and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Transaction on Image Processing*, 2009.

[40] K. Seshadrinathan, R. Soundarajan, A. C. Bovik and L. K. Cormack, "A subjective study to evaluate video quality assessment algorithms," in *SPIE proceding Human Vision and Electronic Imaging*, 2010.

[41] (2010, Mar.) Live Video Quality Database. [Online]. http://live.ece.utexas.edu/research/quality/live

video.html

[42] R. Achanta, F. Estrada, P. Wils, and S. Süsstrunk, "Salient region detection and segmentation," in *International Conference on Computer Vision Systems*, 2008.

[43] X. Hou and L. Zhang, " Saliency detection: A spectral residual approach," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[44] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," *Advances in Neural Information Processing Systems* , vol. 19, pp. 545-552, 2007.

[45] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254-1259, 1998.

[46] Y. F. Ma and H. J. Zhang, "Contrast-based image attention analysis by using fuzzy growing," in *ACM International Conference ACM International Conference*, 2003.

[47] B. S. manjunath, P. Salembier and T. Sikora, *Introduction to MPEG-7 multimedia content description interface*. John Wiley & Sons, LTD, 2003.

[48] K. A. Peker and A. Divakaran, "Framework for measurment of intensity of motion activity of video segments," *Journal of Visual Communication & Image Representation*, vol. 15, pp. 265-284, 2004.

[49] K. A. Peker and A. Divakaran, "A novel pair-wise comparison based analytical framework for automatic measurment of intensity of motion activity of video segments," in *IEEE International Conference on Multimedia and Expo*, 2001.

[50] A. Akutsu, Y. Tonomura, H. Hashimoto and Y. Ohba, "Video indexing using motion vectors," in *SPIE Conference on Visual Communictations and Image Processing*, 1992.

[51] E. Ardizzone, M. laCascia, A. Avanzato and A. Bruna, "Video indexing using MPEG motion compensation vectors," in *IEEE International Conference on Multimedia Computing and Systems*, 1999.

[52] A. Divakaran and K. A. Peker, "Video summarization using motion descriptors," in *SPIE Conference on Storage and Retrieval from Multimedia Database*, 2001.

[53] A. Divakaran, R. Regunathan and K. A. Peker, "Video summarization with motion descriptors," *Journal of Electronic Imaging*, vol. 10, no. 4, 2001.

[54] V. Kobla, D. Doermann, K. I. Lin and C. Faloutsous, "Compressed domain video indexing techniques using DCT and motion vector information in MPEG video," in *SPIE Conference on Storage and Retrival for Image and Video Databases*, 1997.

[55] MPEG-7, "Visual part of the XM 4.0, ISO/iec MPEG99/W3068," 1999.

[56] W. Wolf, "Key frame selection by motion analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1996.

[57] T. Sikora, "The MPEG-7 Visual Standard for Content Description—An Overview," *IEEE Transaction on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 696-702, Jun. 2001.

[58] S. jeannin and A. Divakaran, "MPEG-7 visual descriptors," *IEEE Transaction on Circuits and Systems for Video Technology*, vol. 11, no. 6, 2001.

[59] A. Divakaran, K. A. Peker, H. Sun and A. Vetro, "a supplementary ground-truth dataset for intensity of motion activity," ISO/IEC MPEG00/m5717, 2000.