The role of temporal aspects for quality assessment

Claire Mantel, Thomas Kunlin, Patricia Ladret

To cite this version:

Claire Mantel, Thomas Kunlin, Patricia Ladret. The role of temporal aspects for quality assessment. QoMEX 2010 - Second International Workshop on Quality of Multimedia Experience, Jun 2010, Trondheim, Norway. pp.94-99, 10.1109/QOMEX.2010.5517868 . hal-00536090

HAL Id: hal-00536090

https://hal.archives-ouvertes.fr/hal-00536090

Submitted on 15 Nov 2010

THE ROLE OF TEMPORAL ASPECTS FOR QUALITY ASSESSMENT

Claire Mantel[1,2], *Thomas Kunlin*[1]

[1] STMicroelectronics S.A.
12 Rue Jules Horowitz
B.P. 217
Grenoble - France

Patricia Ladret[2]

[2] GIPSA-Lab - Department of Signal and Images
Grenoble Institute of Technology - UMR CNRS 5216
961 rue de la Houille Blanche
Grenoble - France

ABSTRACT

For quality assessment, videos are often considered as series of images with, at best, a motion component. To study the role of temporal aspects in quality, we compare the perceived quality of two versions of a mosquito noise correction algorithm: one purely spatial and the other spatio-temporal. We set up a paired-comparison experiment specially adapted to the temporal aspects of video quality. Results prove the existence of a purely temporal aspect in video quality perception.

***Index Terms*—** Video quality, subjective experiment, quality assessment, temporal filtering

1. INTRODUCTION

In the field of video quality, the first processings were simply an application of image algorithms to a succession of images. For both artefact reduction and quality assessment, the images were considered independently from one another.
Hamberg et al. showed in [1] that human observers can estimate both instantaneous and continuous quality in a coherent and consistent way, thus proving the importance of measuring quality variations over time.
The first temporal feature added to quality metrics to account for such variations was motion. Almost all metrics work under the assumption that *'the more motion there is, the less noise is perceptible'* but they use different kind of motion information. Localized motion value can be used to weight the spatial quality assessment map for each frame and thus turn them into video quality maps, as done by Li et al. in [2]. The influence of spatial content over temporal artefact can also be taken into account as mentioned by Pinson et al. in [3] through the product of spatial and temporal information (designating the sum of absolute difference between two consecutive frames).
Another issue is the choice of a temporal pooling method: although a simple average over time is the first and most common way, the cognitive mechanisms at work while evaluating video quality are more complex, as shown by Aldridge et al. in [4].
One video quality metric, detailed by Ninassi et al. in [5], not only uses transient and sustained models for temporal perception but also accounts for the temporal variations of spatial artefacts. However, video quality measurement is still mainly approached as a modified image evaluation. In most cases, metrics are composed of successive image quality metric values averaged with a scaling factor over time, and sometimes balanced by a motion quantity coefficient.
The same issue is at stake for subjective assessment methods since the existing methods in the ITU recommendation [6], apart from the SSCQE, are all dedicated to both *'picture and sequence'* as if there was no need to differentiate them. Many issues of subjective assessment methods have been investigated such as the use of a continuous or discrete scale, the experience of subjects in video quality or the influence of methodology over results. Yet there is few documentation about the effect of methodology on temporal defects perception or the evolution of quality through time.
The simple issue of how to turn continuous ratings (SSCQE) into a single rating per sequence (e.g. DSCQS) is still not solved: in [7] Lee et al. average the grades over the whole sequence while in [8] Pinson et al. use only the last rate to represent the complete sequence.
Considering video quality as a 'modified' image quality seems wobbly because it totally disregards the purely temporal aspects of some compression noises and the fact that their temporal evolution is sometimes more annoying than their spatial level. An example of such a noise is the blocking effect on a homogeneous zone. A block switching from one grey level to another is indeed much more noticeable than the same one with constant grey level. Besides, in [9] Itti et al. study which low level features (among color, intensity, orientation, flicker and motion) can predict where an observer gazes in a sequence. Although it does not directly concerns quality assessment, they showed that both motion and flicker were the major gaze attraction factors.
The term 'temporal artefact' is mostly used to name impairments on a sequence timeline, such as frameloss or temporal scaling. This study focuses on temporal compression artefacts and the variation of spatial artefacts through time. Mosquito

noise (MN) occupies a peculiar place in the field of compression noises as it is annoying mainly because of its temporal variation: it is not a major defect for still images. This artefact is located next to the edge of objects, its amplitude is small compared with the grey level variations of edges and varies from one frame to another. For a more complete survey of this noise and the associated correctors, see [10].

We use the corrector explained in the above mentioned paper to investigate the perception of temporal aspects in video quality. We set up a subjective quality assessment experiment, taking great care of enabling observers to assess temporal quality, as detailed in section 2. We had them compare the quality of compressed videos of which correction differs only by the inclusion of a temporal feature described in 3. The video sequences used for the experiment are also meticulously chosen to exhibit mostly temporal artefacts, as explained in section 3. The analysis of the subjective testing results, in section 4, demonstrates the importance of purely temporal aspects for quality perception.

2. METHODOLOGY OF THE EXPERIMENT

We had two objectives in mind while designing our experiment. The first goal was to confront our MN correction algorithm with ground truth. The second was to test whether our spatio-temporal filtering improved the spatial one.

This part describes the overall procedure and the different features of the experiment: the chosen display method, the grading scale, the observers, the debriefing and the set-up used.

2.1. Procedure

The test methodology was not taken 'as is' from the ITU recommendation [6] methodology list because none fitted completely our objectives in terms of display, scale and presentation. It is a combination of features taken from SDSCE and DSIS methods, within the general framework of stimulus comparison method.

Our goal is to evaluate the difference of quality between various versions of video sequences. The variants to compare are the original sequence (of perfect quality), the compressed one and two distinctly corrected sequences: one by the spatial design of our filter and the second by its spatio-temporal version. Those versions are respectively named O, C, S and T in the rest of the paper.

In [7], Lee showed that the presence of a reference video does not have much influence on ratings. To stick the most possible to 'real life' situation we needed to compare extensively each version with all the others and not only with the original one, so we decided not to present the reference video each time and not to identify it.

For each video, we presented all the possible paired combinations to be rated. The observers watched each comparison twice, with a grey screen separation of three seconds in-between. After the second viewing the observers had to answer the question: *'What is the quality difference between the two videos?'*. They answered through a 7-alternative forced-choice method described in section 2.3.

A test session consisted in a training phase of four comparisons, the rating phase and a debriefing with the organizer. During the debriefing, the observers were asked a series of questions about their opinion on the experiment and their rating strategy. The questions asked are stated in Section 2.5 and the answers are mainly discussed in Section 3.

2.2. Display

The ITU recommendation [6] indifferently advises to display the two sequences to compare one after another or simultaneously. The disadvantage of sequential comparison is that the two videos are not directly compared: the memory of the first one is compared with the second. Such a presentation is bound to harm the comparison. Indeed, the quantity of information present in a sequence is too important for our memory to store it all: it is 'coded' with some losses. In [11] Wolfe studies the limits of our visual memory. In particular, they present observers simple synthetic images (several red or green circles on a white background) and they ask them the color of one of those items. After 2 to 12 of those questions, they hide the color of a previously cued item and they ask about it. They show that about 80% of subjects remember well the color of an item they were asked about 2 trials before, but that this rate drops to being not-significantly different from 50% (the hasard rate) for cued items from former trials. If something as simple as the color of a previously cued item leads to that much uncertainty, how can we expect subjects to remember quality information about a 10s video sequence well enough for a comparison?

This observation drove us to present simultaneously the two versions to compare.

Likewise, it seemed to us that the temporal aspect and low intensity of MN would make comparing it precisely enough tough if the sequences are displayed on different screens. That is the reason why we decided to present each pair side-by-side on a single screen, as shown in Figure 1.

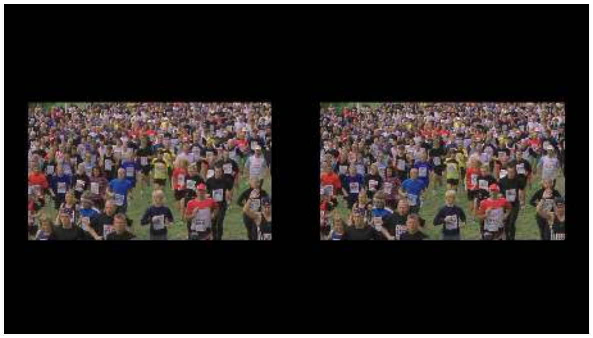


Fig. 1. Display of two CrowdRun versions.

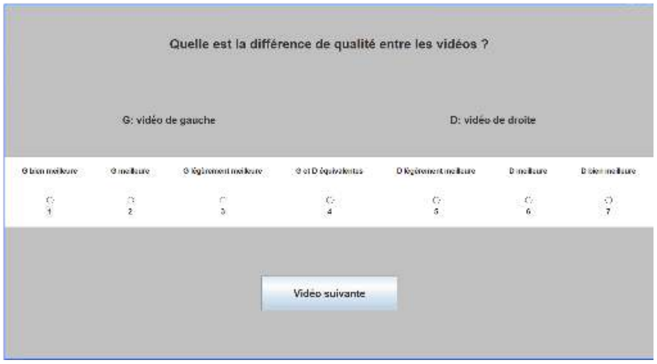
The last question to be addressed for displaying the sequences was their location on the screen. This issue is really important because the drawback of simultaneous presentation is that observers must intermittently examine the two videos. Displaying the two sequences one above the other would be more logical regarding the ocular distance to cover to compare two identical areas. However, Larabi shows in [12] that observers prefer comparing sequences from left to right than from top to bottom so we chose the left/right option.

Every comparison was rated two times during the experiment with the respective location (left or right) of the two versions switching the second time.

2.3. Grading Scale

Studies concerning methodology evaluation have shown that observers do not consider positive and negative affects the same way. In [13], the authors speak of a 'positive-negative asymmetry' consisting in two different mechanisms ('positive bias' and 'negativity effect') that accounts for the difference of judgment expressed for equal negative and positive stimuli. Another issue with an asymmetrical comparison scale is that it establishes a hierarchy between the two videos displayed: one is compared with another, which implicitly makes it a reference for the comparison. A symmetrical scale allows to present them on the same basis.

To prevent those biases, we did not use the ITU comparison scale but a symmetrical one. We kept the number of categories, so the observers had seven different answers possible: three degrees of preference towards the left video *'1 - Left is much better'*, *'2 - Left is better'*, *'3 - Left is slightly better'*, a neutral answer *'4 - Left and Right are equivalent'* and three degrees of preference towards the right video (respectively graded 5, 6 and 7). The notation interface is displayed in Figure 2 (in French).



Quelle est la différence de qualité entre les vidéos ?

G: vidéo de gauche

D: vidéo de droite

G bien meilleure	G meilleure	G légèrement meilleure	G et D équivalentes	D légèrement meilleure	D meilleure	D bien meilleure
○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7

Fig. 2. The notation interface.

2.4. Observers

30 non-paid observers participated, one at a time, in the experiment. They were all naive regarding the purpose of the experiment. All of them are non-experts in video and image processing and all have normal or corrected-to-normal visual acuity and fine color vision (tested with 6 of Ishihara plates). Their ages spread from 24 to 59 years with a mean of 36.9 years and standard deviation of 10. There were 14 women and 16 men.

As having every observer assess each comparison for each video sequence would have been too long, we split the subjects in groups, each one rating a subset of the configuration/video combinations. The rating phase of the experiment was composed of 60 pairs for each subject and it lasted an average of 33 min. Every comparison was rated twice by at least 20 subjects.

2.5. Debriefing

Every observer was asked right after the experiment how he felt about the length of the experiment, the rating scale and the usefulness of the second visualization. He also had to say for each video what drove him to choose his rating: a global impression or some precise elements, and for the later what kind of defects and where he saw them.

All the answers were qualitative and only their appearance frequency was studied.

2.6. Set Up

Observers were placed in front of a Samsung LE40F71B calibrated with a Datacolor Spyder3Elite colorimeter. Concerning the viewing distance, we did not follow the ITU recommendation [14] about paired comparison that advised 8 times the videos height. As our paired videos occupy almost the whole width of our HD TV, we considered that the horizontal ocular angle to be respected was the same as for single video displays and established a distance of three times the height of the TV (48cm).

The room where the experiment took place has light-gray walls, it is isolated from exterior lighting and was lit by two fluorescent daylight-colored (6500K) light bulbs.

3. CHOOSING AND PROCESSING VIDEOS

As we are dealing with a correction algorithm and not a metric, we did not look for extensiveness. To limit the duration of the experiment, the number of sequences used was restricted to 6. Our objective while choosing the videos for the experiment was dual. We wanted sequences that would be the more varied possible while containing mostly temporal artefacts.

The video sequences we used come from the VQEG HDTV and Technical University of Munich databases. As we wanted to display two sequences on one TV screen, we could not use HD format sequences. To avoid any unmastered effect on the video quality, we cropped them rather than downsample them:

first to 720p for encoding and afterwards to another 16:9 format (800x450) for display. Every sequence is cut to last 10s.
Those videos were compressed using the x264 video codec (an open source version of the H264 norm). We encoded each sequence at the bitrate that, in our opinion, created the most temporal artefacts (as defined in section 1) and for which they were the main defects. As the content of the sequences are quite heterogeneous, the chosen bitrates vary widely: from 5,6M for Ducks Take Off to 1,5M for Tractor.

3.1. On the influence of motion on coding quality

Content continuity in video sequences is true in most cases and is for this reason an assumption of the rate control system of encoders. Thus, except after a scene change, they vary slowly enough for us to perceive. As motion information can occupy an important place in the final encoding size (up to 50% for low bitrate videos), the quantity of motion influences greatly the quality of the compressed video.

In this experiment, the necessity for an approximately constant quality during the videos is double. As for any video quality experiment, the recency effect described by Aldridge et al. in [4] informs us that the human visual system is *'quick to criticize, slow to forgive'*, meaning that the assessment of a quality-varying video depends on the moment where the degradations happen.

Moreover, as the type of compression-produced artefacts depends both on the encoder bitrate and the video content, a decrease in motion quantity but not in bitrate might create spatial artefacts in a sequence. As much as rate control systems get better, the sole methods to ensure a constant defect type during a video sequence still are to encode videos with an almost constant encoding difficulty or to chose a different bitrate for each scene. And this is the only way to be sure of what people really rate.

3.2. On the influence of motion on attention

Some preliminary experiment sessions allowed us to see that the previously noted characteristics were not enough to ensure the focus on temporal artifacts.

Since 2005, several labs studying where observers gaze in an image (the locations that are *salient*) have added a temporal feature to their saliency map algorithm to adapt it to videos. In an advanced study on the links between motion and saliency ([9]), Itti has shown that observers are attracted to any object presenting a motion pattern distinct from the global one (that is motion when the camera is still and different-from-background motion when it is moving).

Yet those preliminary sessions confirmed us what Ninassi et al. showed in [15]: ocular behaviors change depending on the task. We had first included the Ice video where ice skaters cross in the picture while a red cone stands still at the center. As we asked the subjects how they rated each video during the post-testing debriefing, everyone of our 6 early observers told us that the cone made him take his decision. This surprising answer was most of the time accompanied by the remark *'it is easier to see defects on something still'*. It is a perfect example of counter natural visual strategy: the video camera is still, there are people moving and yet everyone watches the motionless cone.

This lead us to conclude that to assess temporal quality the whole content of video sequences must be moving.

Another drawback of asking people to rate quality is that if there is a greatly impaired spatial zone (relatively to the rest of the video), once they find it they won't look anywhere else. For example during our preliminary testings, the presence of defects (blocking effect) on a tree occupying maybe a sixth of the screen for 3s in a sequence was all observers watched.

3.3. Scene length

The duration of scenes influences greatly quality rating because it needs to be a conscious process and it is well known that top-down mechanisms are longer than bottom-up ones.

The observers go through several steps to judge quality: they first have to get a global vision and understanding of the scene, then to spot defects and at this point they still need time to compare it willingly with the second video. And this only works for one artefact at a time: as shown in 2.2, our memory is not efficient enough to study several ones simultaneously. Moreover, the specificity of temporal artefacts evaluation is that subjects not only need to see them but also to watch them for a while.

While grading videos, a scene change has the same effect as a perceptual reset button: the content of the screen changes and every landmark disappears.

Two of the videos we used contain a 2s scene. The answer to what the subjects based their rating on for those videos either does not contain any mention of those scenes or subjects said straight that they *'did not see anything'* during this scene.

For those reasons, video sequences selected for quality assessment should contain a single scene or at least the scenes length should be more than 2s.

3.4. Features of the algorithm used

To study the impact of temporal continuity over video quality, we used two versions of the corrector presented in [10]. They both use the same Variation-Inverse Filter but the chosen support differs. The first variant, called the spatial version, uses pixels present in a 3-by-3 neighborhood of the current pixel. Whereas the second sort, the spatio-temporal version, extends the spatial neighborhood with pixels selected among the previous and the following frame according to a 'belonging to the same object' criteria. For us knowing if the pixels moved from one frame to another is not relevant, the only requirement to append pixels to the support is whether they are part of the same object.

4. RESULTS

4.1. Statistical analysis tools and terminology

We used the method recommended in [6] and [16] to detect outlier observers but none got rejected.

To analyze the results of our experiment, we chose to apply classical experimental psychology methods. To estimate the accuracy of each of our research hypotheses, we try to reject the opposite assumption: the *null hypothesis*.

To this aim, we used a very common hypothesis test: the Student t-test. As explained in [17], this robust test assesses the significance of the mean value of samples. Here we apply this test in two different cases:

- to estimate the probability that the obtained samples from one comparison are extracted from a distribution of average 4. Indeed, as this value is the center of our scale and means that the two compared versions are similar, we need to know if the mean difference with this centered value is significant. C-S, C-T and T-S comparisons are respectively analyzed this way in sections 4.3.1, 4.3.2 and 4.4.

- to assess the probability that the mean difference between two comparisons is significant and cannot derive from our samples ('paired t-test'). The difference between C-S and C-T results is thus characterized in 4.4.

We set a threshold at 0.05, signifying that for a t-test result (designated by 'p') below this value, we will consider our research hypothesis validated. The lower the p value, the better the result. In this case, the t-test actually informs us that there is more than 95% chance that it is true.

We also studied the effect of the grade rank (first or second), the location on the screen (left or right) and the gender of observers on their grading. There is no significant influence of any of those features on the obtained results.

4.2. Protocol validation

The first thing to assess is whether the observers clearly perceived the deteriorations done to the original version. To do so, we use the cumulative relative frequency chart for O-C, O-S and O-T comparisons. It represents the proportion of scores for a category and the preceding ones. For the three comparisons more than 72% of observers judged the original version of the sequence *'better'* or *'much better'*, and more than 90% of them at least *'slightly better'*. There is no questioning that the original version perception stands out from the others.

We also investigated the difference between O-C, O-S and O-T grades but there is no significant variation. It means that the quality of the original version is so much better than the others that they seem all leveled when compared with it.

4.3. Algorithm validation

Figure 3 displays the subjective testing results averaged over the two notes and all the subjects with the associated confidence intervals. When one of those averages per comparison is below the central value (4), it means that the first element of the comparison was graded better than the second. For example the O-C comparison bar indicates that the original version (O) was clearly preferred to the compressed one (C). Note that in this figure the ordinates range from 1 to 5 (instead of 1 to 7) since no average reaches a higher value. To analyze the remaining results, we used those averages and t-tests to state if the difference to the central value is significant.

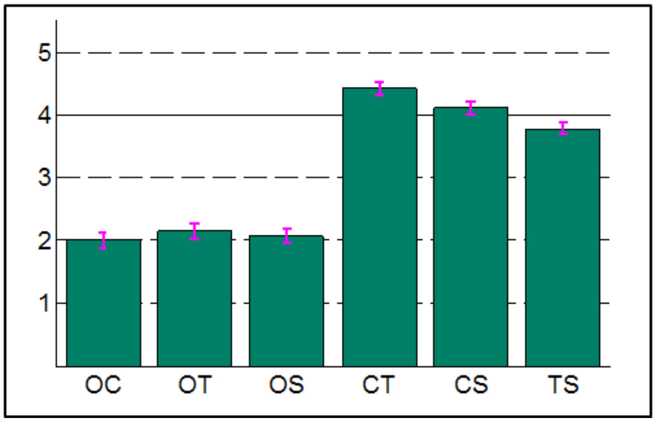


Fig. 3. Average and confidence interval of grades for each comparison

4.3.1. Spatial version

We study here the perceived difference between the compressed sequences and the spatially-corrected ones, that is the CS bar in Figure 3. The corresponding t-test result is $p < 0.05$, meaning that the obtained ratings are significantly different from the central value of the scale. This confirms that the spatial-only filter enhance visual quality.

4.3.2. Spatio-temporal version

The efficiency of the temporal correction is visible through its comparison with the compressed version: the CT comparison in Figure 3. We obtain a t-test value of $p < 0.001$, indicating that observers significantly thought that the spatio-temporal correction improved the sequence quality.

4.4. Spatial versus spatio-temporal processing

There are two different ways to study the quality difference between our two corrections. First, we can analyze the results of the direct comparison between the spatial and spatio-temporal correction: the TS bar in Figure 3. The related t-test value is $p < 0.001$: when their eyes are set on both corrections, subjects significantly prefer the spatio-temporal one.

The second option is the indirect comparison through the compressed version. Indeed, the interrogation we wanted to answer can be stated: *'what correction improves a compressed sequence the best?'*. To do so, we investigate if the average difference between the CS and CT comparisons are significant with a paired t-test. The result is also positive ($p < 0.05$), meaning that the difference between spatio-temporal and compressed versions is more important than between the spatial and compressed versions.

Anyway we analyze results, they establish that observers prefer the spatio-temporal correction to the spatial one. As those corrections differ only by the temporal aspect, this proves the existence of a purely temporal part in video quality.

5. CONCLUSION

Our goal was to study the importance of the temporal properties of video quality through the application of a MN algorithm on a spatial and on a spatio-temporal support. We set-up a subjective paired-comparison experiment to obtain ground truth on the relative quality of four versions of videos: perfect, compressed and processed with both processings. Several aspects of the experiment methodology are specified to account for the temporal singularity of video quality assessment. The analysis of subjective quality ratings demonstrates the preference of observers for the spatio-temporal version of the algorithm over the purely spatial one.

Those results ascertain the reality of a purely temporal part in video quality. They also validate the interest of taking temporal specificities into account while designing quality assessment methodologies.

Now that the perception of temporal continuity and its role as to video quality are established, the next logical step is to see how the current quality metrics account for this phenomenon. To study this connection, we will confront those results with various video quality metrics.

6. REFERENCES

[1] R. Hamberg and H. de Ridder, "Continuous assessment of perceptual image quality," *Journal of the Optical Society of America A*, vol. 12, pp. 2573–2577, 1995.

[2] Q. Li and Z. Wang, "Video quality assessment by incorporating a motion perception model," in *International Conference on Image Processing, ICIP*, 2007, pp. 173–176.

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

[4] R.P. Aldridge, D.S. Hands, D.E. Pearson, and N.K. Lodge, "Continuous quality assessment of digitally-coded television pictures," *Vision Image and Signal Processing, IEE Proceedings*, vol. 145, pp. 116–123, 1998.

[5] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, "Considering temporal variations of spatial visual distortions in video quality assessment," *IEEE Journal Of Selected Topics In Signal Processing : Special Issue On Visual Media Quality Assessment*, vol. 3, no. 2, pp. 253–265, 2009.

[6] ITU-R, *Recommandation BT.500-11 Methodology for the subjective assessment of the quality of television pictures*, 2002.

[7] C. Lee, H. Choi, E. Lee, S. Lee, and J. Choe, "Comparison of various subjective video quality assessment methods," in *SPIE Electronic Imaging*, 2006, vol. 6059.

[8] M. Pinson and S. Wolf, "Comparing subjective video quality testing methodologies," *SPIE Electronic Imaging*, vol. 5150, pp. 8–11, 2003.

[9] L. Itti, "Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes," *Visual Cognition*, vol. 12, no. 6, pp. 1093–1123, 2005.

[10] C. Mantel, P. Ladret, and T. Kunlin, "A temporal mosquito noise corrector," in *Quality of Multimedia Experience, QoMEx 2009. International Workshop on*, 2009, pp. 244–249.

[11] J. M. Wolfe, A. Reinecke, and P. Brawn, "Why don't we see changes? the role of attentional bottlenecks and limited visual memory," *Visual Cognition*, vol. 14, no. 4-8, pp. 749 – 780, 2006.

[12] M.-C. Larabi, "Comparison of subjective assessment protocols for digital cinema applications," *SPIE Electronic Imaging*, vol. 7529, pp. pp1–10, 2010.

[13] G. Peeters and J. Czapinski, "Positive-negative asymmetry in evaluations: The distinction between affective and informational negativity effects," *European Review of Social Psychology*, vol. 1, pp. 33–60, 1990.

[14] ITU-T, *Recommandation P910. Subjective video quality assessment methods for multimedia applications*, 2008.

[15] A. Ninassi, O. LeMeur, P. Le Callet, D. Barba, and A. Tirel, "Task impact on the visual attemtion in subjective image quality assessment," *European Signal Processing Conference, Eusipco*, vol. 5, pp. 802–817, 2006.

[16] VQEG, *Test Plan for Evaluation of Video Quality Models for Use with High Definition TV Content v3.0*, 2009.

[17] D. C. Howell, *Statistical methods for psychology, 7th Edition*, Wadsworth, 2010.