Adaptive testing for video quality assessment

Adaptive testing for video quality assessment

Vlado Menkovski
Eindhoven University of Technology
Postbus 513
5600 MB Eindhoven
+31 402475653
v.menkovski@tue.nl

Georgios Exarchakos
Eindhoven University of Technology
Postbus 513
5600 MB Eindhoven
+31 402475653
g.exarchakos@tue.nl

Antonio Liotta
Eindhoven University of Technology
Postbus 513
5600 MB Eindhoven
+31 402473890
a.liotta@tue.nl

ABSTRACT

Optimizing the Quality of Experience and avoiding under or over provisioning in video delivery services requires understanding of how different resources affect the perceived quality. The utility of resources, such as bit-rate, is directly calculated by proportioning the improvement in quality over the increase in costs. However, perception of quality in video is subjective and, hence, difficult and costly to directly estimate with the commonly used rating methods. Two-alternative-forced choice methods such as Maximum Likelihood Difference Scaling (MLDS) introduces less biases and variability, but only deliver estimates for relative difference in quality rather than absolute rating. Nevertheless, this information is sufficient for calculating the utility of the resource on the video quality. In this work, we are presenting an adaptive MLDS method, which incorporates an active test selection scheme that improves the convergence rate and decreases the need for executing the full range of tests.

General Terms

Maximum Likelihood Difference Scaling, adaptive MLDS, Video Quality Assessment (VQA), Quality of Experience, QoE.

1. INTRODUCTION

The goal of efficient management for video delivery services is delivering the desired Quality of Experience (QoE) without over-provisioning the service resources. To make this process feasible we need an understanding of the relationship between the resources and the delivered quality. Moreover, if we can measure the utility of each resource, such as bit-rate, for the perceived quality, we can then provide this resource optimally or up to the level that is justified by the cost. For example, depending on the context, type of content and screen characteristics a person might not perceive any more improvement if the video bit-rate is larger than 512kbps. On the other hand, for a low cost service a 256kbps video could offer only slightly lower quality than 512kbps (again in the specific context) and be the optimal option. Calculating these utilities requires understanding of the costs, but more importantly, it requires understanding of the perceived quality for these resources.

Measuring the relationship between a resource provided by the video delivery service and the provided quality requires subjective testing because of the subjective nature of perceived quality of video. Objective and subjective video quality methods have varied levels of success in delivering accurate estimations. The objective methods are considered more practical because they do not necessitate human testing. Nevertheless, they are less accurate mainly because they do not consider all the factors that affect the quality and disregard the viewers' expectations [1].

The subjective methods are regarded as more accurate and are usually used as a benchmark for the objective methods. One such study by Seshadrinathan et al. [2] analyzes the different objective video quality assessment algorithms by correlating their output with the differential mean opinion score (DMOS) of a subjective study they executed. This type of undertaking is costly, time consuming and necessitates considerable amount of tests to achieve statistical significance. The bias and the variability of subjective testing arise from the fact that subjective tests rely in rating as the estimation procedure. Rating is inheritably biased due to the variance in the internal representation of the rating scale by the subjects [3][4][5]. Another subjective testing method uses the scale of just noticeable differences (JND). The JND scale measures the amount of subjective impairment in the video. One unit of JND corresponds to the amount of difference that is just noticeable (usually 50% of the time) and as such spans through the whole range of the physical parameter of interest [6]. However, this method requires multiple iterations through different levels of stimuli intensity to determine the scale of JND and cannot directly scale for example a given video with 10 arbitrary levels of bit-rate. Additionally, the JND unit will not be constant on a wider range of bit-rates, which are of interest in practical cases.

In our previous work [7] we have used a two-alternative-forced-choice (2AFC) method to estimate the relative differences in quality. The method Maximum Likelihood Difference Scaling (MLDS) delivers the ratio of subjective quality between a video with different levels of resource provided. Because the method is 2AFC, meaning the participant is forced to choose between two intensities, the amount of bias and variability is significantly lower than in rating [8]. In the case of video quality estimation the 2AFC test is discriminating between different levels of quality. Four videos or two pairs of video are presented and the respondent needs to select which pair has the bigger difference in quality. This might sound as a particularly difficult and time-consuming effort, but in reality most of the tests are quickly and easily answered. The video is typically short (less than 10 seconds) and uniformly impaired, so in most cases the participant is confident enough to vote after only watching a part of each of the video. Many of the tests are quite obvious and derivative, i.e. based on previous responses the following are apparent. Nevertheless, if one wants to explore additional parameters, such as the type of video or the context in which its being watched the number of tests increases quickly. For example in the study executed in [7] for 10 types of videos a participant needed to answer 210 tests per video. Answering all the 2100 tests for each participant took around 8h over the period of a week.

Motivated by the effectiveness of MLDS in estimating the utility of the resources for video quality and its drawback in the number

of tests that quickly grows with the number of samples and parameters under test we have developed an active testing procedure adaptive MLDS. This approach leads to significant decrease in the number of tests and improvement in the learning rate.

2. MLDS

The goal of the MLDS method is to map the objectively measurable scale of video quality to the internal psychological scale of the viewers. The output is a quantitative model for this relationship based on a psychometric function [9] as depicted in Figure 1.

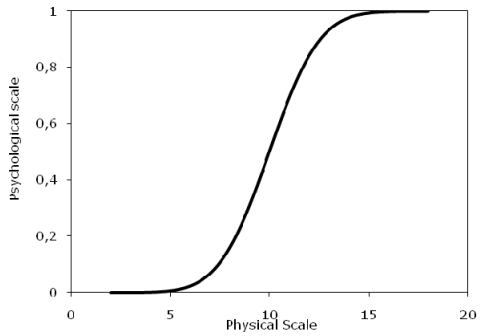

Figure 1. Psychometric function.

The horizontal axis of the Figure 1 represents the physical intensity of the stimuli – in our study the bit-rate of the video. The vertical axis represents the psychological scale of the perceived difference in quality. The perceptual difference of quality ψ_1 of the first (or reference) sample x_1 is fixed to 0 and difference of quality ψ_{10} of the last sample x_{10} is fixed to 1 without any loss in generality [10]. In other words, there is 0% difference in quality from x_1 to x_1 (itself), while there is 100% difference in quality from x_1 to x_{10}. The MLDS method estimates the relative distances of the rest of the videos ψ_2 through ψ_9 and therefore models the viewers' internal quality scale.

This 2AFC test is designed in the following manner; two pairs of videos are presented to the viewers $\{x_i, x_j\}$ and $\{x_k, x_l\}$ where the indexes of the samples are selected as $1 \le i < j < k < l \le 10$, so that the ranges of quality does overlap. The video with smaller index has higher quality. The viewer then selects the pair of videos that have bigger difference in quality. For a given test T_n the viewer selects the first pair (sets $R_n=1$) if she perceived the qualities of videos in the quadruple as $|\psi_j - \psi_i| - |\psi_l - \psi_k| > 0$, otherwise she chooses the second pair ($R_n=0$). These comparisons between the quality distances of video pairs allow for design of a quality distance model between all of the presented videos. The method calculates the quality differences ψ_2 through ψ_9 as parameters in maximum likelihood estimation (MLE).

The MLE requires a probability distribution for each response. This is done using signal detection theory (SDT). The difference of differences of quality between the four videos is the signal contaminated by Gaussian noise. When executing a test the participant calculates the value $\delta(i,j,k,l)_n + \varepsilon = \psi_{j_n} - \psi_{i_n} - \psi_{l_n} + \psi_{k_n} + \varepsilon$ where ε is value sampled from a Gaussian distribution with zero mean and standard deviation of 1.

Using this assumption, the probability of each response is $P(R_n = 1; \delta_n, \sigma^2) = 1 - \Phi\left(\frac{0 - \delta_n}{\sigma}\right) = \Phi(\delta_n)$ for a test where the first pair is selected and $P(R_n = 0; \delta_n) = 1 - P(R_n = 1; \delta_n) = 1 - \Phi(\delta_n)$ for a test where the second pair is selected. The likelihood of all the responses is: $L(\Psi \mid \vec{R}) = \prod_{n=1}^{N} \Phi(\delta_n)^{R_n} \left(1 - \Phi(\delta_n)\right)^{1-R_n}$. There is no closed form for such a solution, so a direct numerical maximization method needs to be used to compute the estimates $\hat{\Psi} = \arg\max_{\Psi} L(\Psi \mid \vec{R})$. More details on MLDS for video quality can be found in [7] and for image quality in [11].

A fitter curve through the $\hat{\Psi}$ also represents the utility of the bit-rate as a resource or how much we can improve the quality by increasing the bit-rate over the tested range assuming that the cost of increasing the bit-rate is constant over the same range.

3. Adaptive MLDS

The MLDS method is appealing for its simplicity and efficiency, however one full round of tests for ten levels of stimuli (i.e. video qualities) requires 210 individual tests. The full range of tests carry significant redundancy and removing some of it should not necessarily make the results significantly less reliable; even more so it can have only negligible effects on the end result.

In this adaptive procedure we have two aims, to improve the rate of learning and to decrease the number of required tests. The approach is based on the idea that with the knowledge acquired by executing a small number of tests we can estimate the answers of the remaining tests with some confidence. Then using these estimates together with the known responses we execute the MLDS method. Executing the MLDS with more responses helps the argument maximization procedure. The estimates rely on the characteristics of the psychometric curve (such as its increasing monotonicity), so that the overall performance of MLDS is improved. The idea comes from the notion that some of the tests are covering the range of others. In fact, all of the tests are being covered by others in one way or the other. The approach makes use of the characteristics of the psychometric curve. The psychometric curve is a monotonously increasing function $\vec{\Psi} = f(\vec{X})$. Consequently, for $k < l < m$, $x_k > x_l > x_m$ if $x_k - x_l > x_k - x_m$ in the physical domain then $\psi_k - \psi_l \ge \psi_k - \psi_m$ in the psychological domain Figure 2.

If we now observe five samples x_i, x_j, x_k, x_l, x_m such that $i < j < k < l < m$ and we observe two tests $T_1(x_i, x_j; x_k, x_l)$ and $T_2(x_i, x_j; x_k, x_m)$, the perceived qualities in the psychological domain are $\psi_i \le \psi_j \le \psi_k \le \psi_l \le \psi_m$. If in T_2 the first pair is bigger or $\psi_j - \psi_i > \psi_m - \psi_k$ that would mean that $\psi_j - \psi_i > \psi_m - \psi_k \ge \psi_l - \psi_k$. In other words, if in T_2 the first pair is selected with a bigger difference, then in T_1 the first pair has a bigger difference as well (Figure 2).

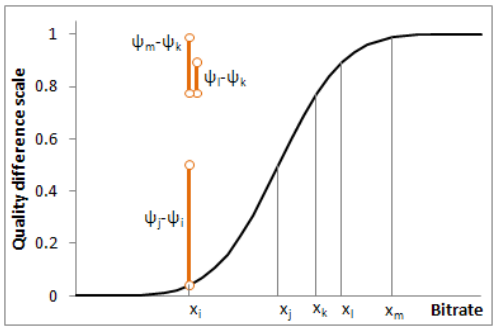

Figure 2. Monotonicity of the psychometric curve

There are many different combinations of tests that have this dependency for the first pair or the second pair. We can generate a list of dependencies for each pair based on two simple rules:

- Let us assume test $T_1(a, b, c, d)$ such that $a < b < c < d$, $\psi_b - \psi_a > \psi_d - \psi_c$ and test $T_2(e, f, g, h)$ with $e < f < g < h$. If $e \le a < b \le f$ and $c \le g < h \le d$ then $\psi_f - \psi_e > \psi_h - \psi_g$.
- Let us assume that for test $T_1(a, b, c, d)$ with $a < b < c < d$, $\psi_b - \psi_a < \psi_d - \psi_c$. If for test $T_2(e, f, g, h)$ with $e < f < g < h$ the following hold: $a \le e < f \le b$ and $g \le c < d \le h$ then $\psi_f - \psi_e < \psi_h - \psi_g$.

After introducing an initial set of responses we can estimate the probabilities of the rest, however first we need to learn the probabilities of each of the known responses to be actually valid. MLDS estimates the values of the psychological parameters $\Psi=(\psi_1,...,\psi_{10})$ such that the combined probabilities of each response or the overall likelihood of the dataset is maximized. Nevertheless, after the argument maximization is finished the different responses have different probabilities of being true.

Having a set of initial quality Ψ values as the prior knowledge about the underlying process coming from the data, we generate the estimations for the rest of the tests. The interdependencies from the tests are far more complex, of course.

Let us assume, for example, a test T_1 that depends on tests T_2 and T_3. If the answer from T_2 indicates that the first pair has a larger difference in T_1 and the answer from T_3 indicates the opposite then we need to calculate the combined probability of T_2 and T_3 to estimate the answer of T_1.

Assuming that the responses of T_2 and T_3 are independent and that the probability of giving the first and second answer is the same, the combined probability of T_2 and T_3 is

$$P(T_1) = \frac{P(T_2)(1 - P(T_3))}{P(T_2)(1 - P(T_3)) + (1 - P(T_2))P(T_3)}.$$

Of the remaining tests that have no responses, some will have higher estimates than others. In other words we have better estimations for some of tests than others. To improve the speed of learning, the adaptive MLDS method, focuses on tests that have smaller confidence in the estimations. This way when we receive the next batch of responses the overall uncertainty in the estimates should be minimized.

The goal of the adaptive MLDS is to develop a metric that will indicate how sufficient the amount of tests is for determining the psychometric curve. We can obtain this indication from the probabilities of the estimations. As we get more responses by asking the right questions the estimation for the rest of the tests improves. At some point adaptive MLDS will have very high probabilities of estimating correctly all of the remaining tests. This is a good indication that no more tests are necessary.

4. EXPERIMENTAL SETUP

To show the performance of the adaptive MLDS we have developed a software simulation, which simulates the learning process of the adaptive MLDS algorithm by sequentially introducing data from the subjective study in [7]. For every iteration a psychometric curve is estimated and compared to the one calculated on the full dataset. The root mean square error (RMSE) is computed on the differences. In parallel a random introduction of data is also executed as a baseline for comparison. The adaptive MLDS algorithm is implemented in Java, while the MLDS software from [10] written in R is used for estimating the psychometric curves.

5. RESULTS

Adaptive MLDS as an active learning algorithm explores the space of all possible 2AFC tests with the goal of optimizing the learning process. It also provides indication of confidence in the model built on the subset of the data, so that early stopping of the experiment is feasible. The performance of the adaptive MLDS is presented in Figure 3, 4 and 5. In Figure 3 we present the accuracy of the estimations for three types of videos (blue sky, sun flower and mobile & calendar) against the number of introduced datapoints. In Figure 4, we observe the leaning rate of adaptive MLDS against the random MLDS. The horizontal axis represents the number of points introduced at the time the calculation was executed and the vertical axis the RMSE between the estimated curve and the curve built on the whole dataset. We can clearly observe that for this datapoints adaptive MLDS brings significant improvement in the learning rate. The experiment was repeated for 100 times for each number of datapoints introduced starting from a different random 15 datapoints. In Figure 5 we present the standard deviation of the different value for the RMSE at each point. Figure 6 presents the distribution of the confidence or the probabilities of those estimations. The data in Figure 6 shows that the adaptive learning algorithm estimated the unknown answers with high confidence and that after between 40 and 60 collected answers the confidence in the estimations was close to 1, suggesting that the rest of the tests are not necessary and that we can correctly estimate the psychometric curve without them. This also evident in Figure 3 where the accuracy surpasses 94-96% after 60 tests.

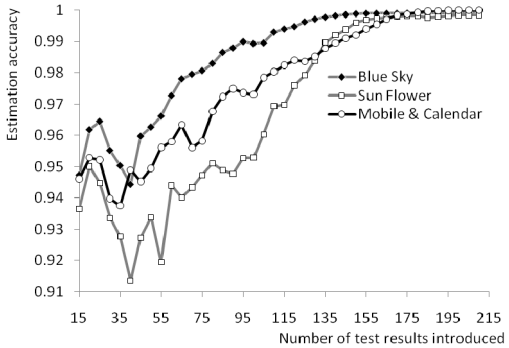

Figure 3. Accuracy of the estimations.

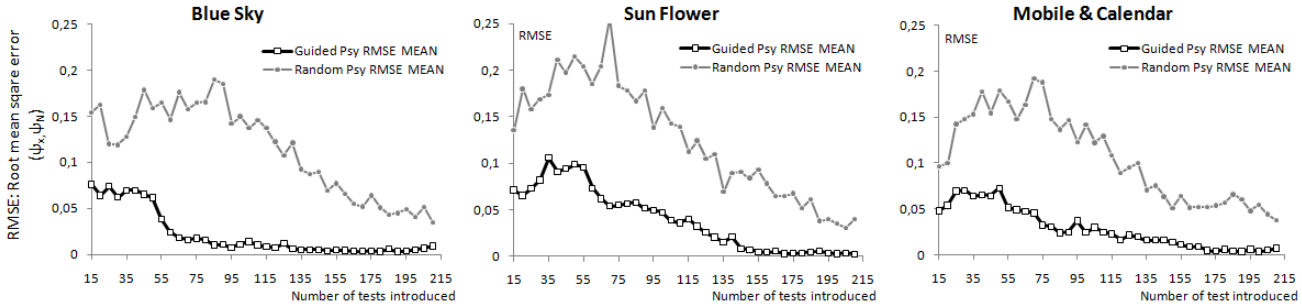

Figure 4. Mean RMSE for the three types of video

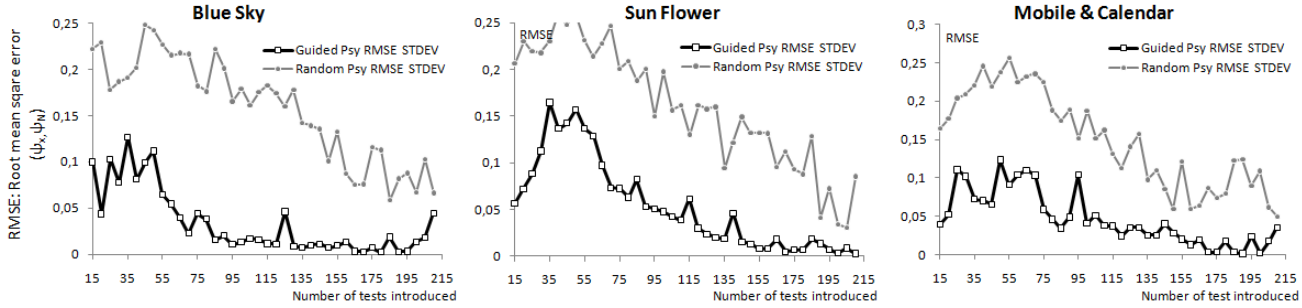

Figure 5. Standard deviation of the RMSE for the three types of video

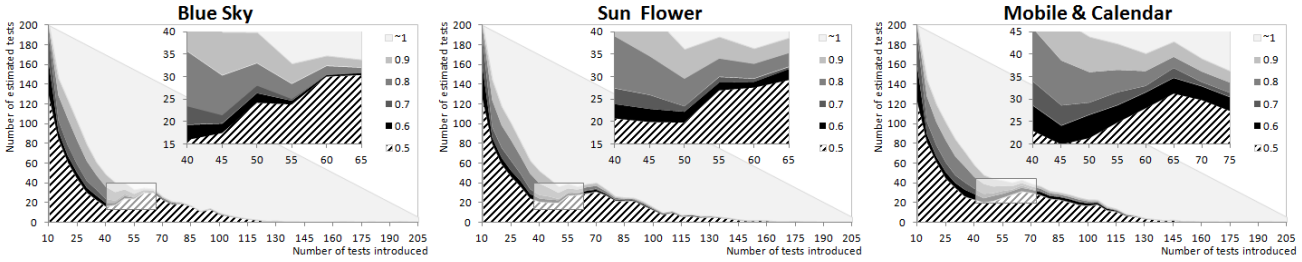

Figure 6. Estimation confidences for the tree types of videos over the number of introduced datapoints

6. CONCLUSION

The adaptive MLDS algorithm is an active learning algorithm specifically designed for the MLDS method for estimating a psychometric curve. Motivated by the fact that MLDS is efficient in estimating video quality utility functions we have developed this adaptive scheme to improve the learning efficiency. The results from the simulations show that adaptive learning provides for significant improvement in the learning rate of MLDS and gives solid indication for stopping the test early when further tests bring no significant improvement in the accuracy of the psychometric curve. Overall this approach adds to the efficiency of MLDS into tackling the issues that arise with subjective estimations of video quality.

7. REFERENCES

[1] S. Winkler and P. Mohandas, "The Evolution of Video Quality Measurement: From PSNR to Hybrid Metrics," *Broadcasting, IEEE Transactions on*, vol. 54, no. 3, pp. 660-668, 2008.

[2] A. Kalpana Seshadrinathan, B. Rajiv Soundararajan, C. B. B. Alan, and K. C. B. Lawrence, "A Subjective Study to Evaluate Video Quality Assessment Algorithms."

[3] D. H. Krantz, R. D. Luce, P. Suppes, and A. Tversky, "Foundations of measurement, vol. 1: Additive and polynomial representations," *New York: Academic*, 1971.

[4] R. N. Shepard, "On the status of'direct'psychophysical measurement," *Minnesota studies in the philosophy of science*, vol. 9, p. 441–490.

[5] R. N. Shepard, "Psychological relations and psychophysical scales: On the status of," *Journal of Mathematical Psychology*, vol. 24, no. 1, p. 21–57, 1981.

[6] A. B. W. A and L. K. B, "Measurement of visual impairment scales for digital video."

[7] V. Menkovski, G. Exarchakos, and A. Liotta, *The value of relative quality in video delivery*. Eindhoven, Netherlands: Eindhoven University of Technology, 2011.

[8] A. B. Watson, "Proposal: Measurement of a JND scale for video quality," *IEEE G-2.1. 6 Subcommittee on Video Compression Measurements*, 2000.

[9] W. H. Ehrenstein and A. Ehrenstein, "Psychophysical methods," *Modern techniques in neuroscience research*, p. 1211–1241, 1999.

[10] K. Knoblauch and L. T. Maloney, "MLDS: Maximum likelihood difference scaling in R," *Journal of Statistical Software*, vol. 25, no. 2, p. 1–26, 2008.

[11] C. Charrier, L. T. Maloney, H. Cherifi, and K. Knoblauch, "Maximum likelihood difference scaling of image quality in compression-degraded images," *Journal of the Optical Society of America A*, vol. 24, no. 11, p. 3418–3426, 2007.