Tackling the Sheer Scale of Subjective QoE

Vlado Menkovski, Georgios Exarchakos, and Antonio Liotta

Eindhoven University of Technology,
P.O. Box 513, 5600MB Eindhoven, The Netherlands
{v.menkovski,g.exarchakos,a.liotta}@tue.nl

Abstract. Maximum Likelihood Difference Scaling (MLDS) used as a method for subjective assessment of video quality alleviates the inconveniencies associated with high variation and biases common in rating methods. However, the number of tests in a MLDS study rises fairly quickly with the number of samples that we want to test. This makes the MLDS studies not scalable for the diverse video delivery environments commonly met in pervasive media networks. To tackle this issue we have developed an active learning approach that decreases the number of MLDS tests and improves the scalability of this method.

Keywords: Maximum Likelihood Difference Scaling, adaptive MLDS, Video Quality Assessment, Quality of Experience, QoE.

1 Introduction

As video is becoming highly pervasive, pervasive media networks are being developed as an underlying delivery technology to handle the newly arisen technical requirements. Pervasive media networks deliver and adapt video and other multimedia content to the context, environment and purpose for which the content is being used. Efficient adaptation of the different video parameters necessitates understanding of the effect of these parameters on the delivered Quality of Experience (QoE). 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 setting. Calculating these utilities requires understanding of the costs, but more importantly it requires understanding of the perceived quality for these resources. To determine the utility of these resources an accurate estimation of quality is necessary. This needs to be achieved through subjective testing, because of the subjective nature of perceived quality of video.

Our focus is on Maximum Likelihood Difference Scaling (MLDS) because of its superior performance as a subjective testing methodology. MLDS is based on two-alternative-forced choice (2AFC) tests that suffer significantly less from bias and variability [1]. However, MLDS studies require all the combinations of four for a given set of samples. As the number of parameter or characteristic of interest

L. Atzori, J. Delgado, and D. Giusto (Eds.): MOBIMEDIA 2011, LNICST 79, pp. 1–15, 2012.

increases, so does the number of samples and in turn the number of MLDS tests. Even though each of these 2AFC tests is simple and straight forward the overall subjective study is not scalable. To tackle the scale of this type of subjective studies we have developed an adaptive test selection procedure for MLDS that improves the learning rate. The adaptive approach iteratively inputs new data by asking the participant to do specific tests, instead of randomly going through all of the combinations of samples. Because of the built in redundancy and high correlation in MLDS, some tests become more informative than others over the course of the experiment. The adaptive MLDS estimates the responses of the unknown tests from the information collected by the answered ones. The tests estimated with less confidence are more informative and are selected as next. Additionally, the confidence for the remaining unknown tests is an indication of how much more tests are necessary, and provides for early stopping capability.

The adaptive MLDS algorithm implemented in a software test bed and executed over subjective test data showed significant improvement in the learning rate and substantial decrease in the number of tests that are necessary.

2 Video Quality Assessment

Estimation of video quality is highly diverse area with many methods, which fall within the two main categories of objective or subjective. Objective methods estimate the quality by focusing on the signal fidelity or measuring the distortions of the video compared to the original. These objective methods are referred to as full-reference (FR) methods. Some effective FR methods include MultiScale-Structural SIMilarity index (MS-SSIM) [2], Perceptual Video Quality Metric (PVQM) [3], and the perceptual spatio-temporal frequency-domain based MOtion-based Video Integrity Evaluation (MOVIE) [4]. These vary in accuracy compared to the subjective reference estimation and complexity. In these methods there is always a trade-off between accuracy and complexity and memory requirements. In addition to the FR methods there are the reduced-reference (RR) and no-reference (NR) objective methods. The RR methods have only partial information on the original signal. Although less accurate this makes them more practical than FR and applicable to continuous assessment of video quality while FR are mostly used in offline estimation. One such method is [5], which examines local harmonic strength features. These features are correspond to artifacts such as blockiness and blurriness. By examining the loss of these features the method estimate the video quality. The NR methods are the most practical because the hold no information on the source of the signal, but also most challenging to implement.

The subjective methods include some type of tests with actual human participants. Evidently objective tests are more practical and therefore with significantly more widespread use. However, objective tests commonly are not designed to consider all the factors that affect the perceived quality of the video or the QoE [6]. In this manner 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. [7] 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 [8][9][10].

In [8] we describe the use of 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 a 2AFC method, meaning the participant has to answer a single binary question of the 'which is bigger' type, the amount of bias and variability is significantly lower than in rating [1].

In the case of video quality estimation the 2AFC test is discriminating between different levels of quality. More particularly, four videos or two pairs of video are presented and the participant needs to select which pair has the bigger difference in quality. This might sounds as a particularly difficult and time-consuming effort, but in most cases the difference in video quality is quite evident. The video is typically short (less than 10 seconds) and uniformly impaired, so very often 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, the number of tests is combination of all the samples over four, so the number of tests is a function that is forth order polynomial of the number of samples. For one or two parameters that affect the video the number of samples is not very big, but as number of samples grows the tests become unfeasible.

Motivated by the effectiveness of MLDS in estimating difference in quality or the utility of the resources and the possibilities for improving the efficiency of the method we have developed an adaptive test selection procedure for MLDS that improves the learning rate and provides for possibilities for executing a subset of the subjective tests while estimating the rest with a given confidence.

2.1 MLDS

To better understand the mechanics of the adaptive MLDS we need to start with a discussion on MLDS itself. The goal of this 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 [11] as depicted in Figure 1.

The horizontal axis of the Figure 2 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 [12]. In other words there is 0% difference in quality between x_1 and x_1 (itself), while there is 100% difference in quality between x_1 and 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.

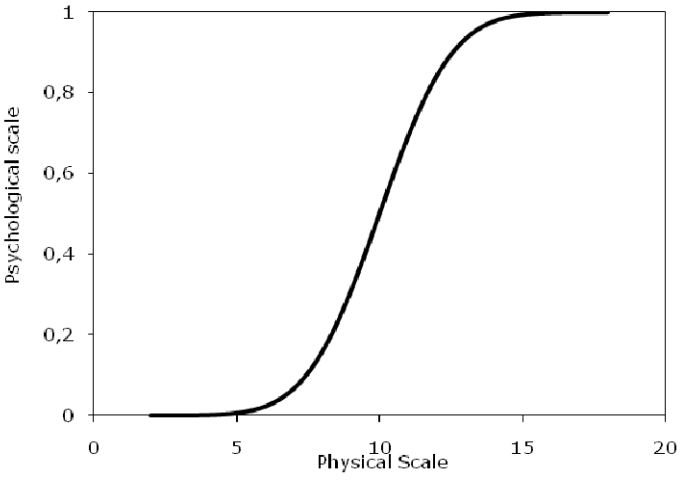

Fig. 1. Psychometric function

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 $1 \leq i < j < k < l \leq 10$, 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 is a method for estimating the parameters of a statistical model. Using signal detection theory (SDT) [13] MLDS models each response as sampled from a Gaussian distribution with unknown parameters. The difference of differences of quality between the four videos is the signal contaminated by Gaussian noise or the mean of a Gaussian distribution (1). When executing a test the participant calculates the value.

$$\delta(i,j,k,l)_n + \varepsilon = \psi_{j_n} - \psi_{i_n} - \psi_{l_n} + \psi_{k_n} + \varepsilon \tag{1}$$

Where ε is value sampled from a Gaussian distribution with zero mean and standard deviation of 1.

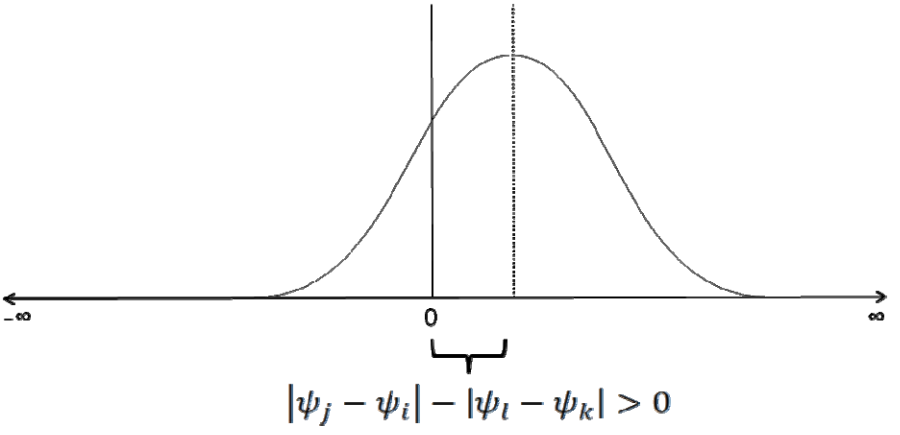

Fig. 2. Stimuli intensity contaminated with Gaussian noise

Using this assumption, the probability of each response is given in (2).

$$P(R_n = 1; \delta_n, \sigma^2) = 1 - \Phi\left(\frac{0 - \delta_n}{\sigma}\right) = \Phi\left(\delta_n\right) \tag{2}$$

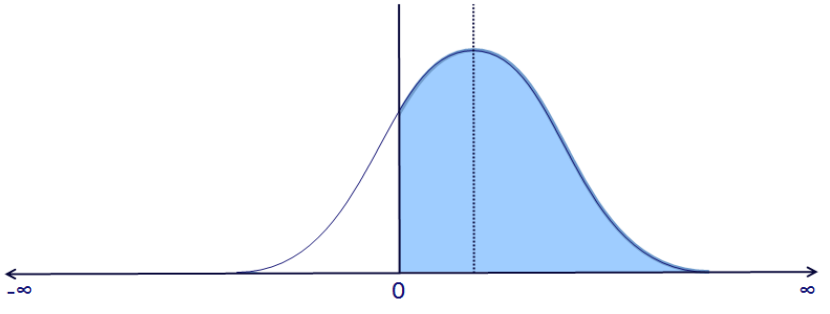

Fig. 3. Probability of first pair being selected

For a test where the first pair is selected the probability is given in (2).

$$P(R_n = 0; \delta_n) = 1 - P(R_n = 1; \delta_n) = 1 - \Phi\left(\delta_n\right) \tag{3}$$

For a test where the second pair is selected the probability is given in (3). The likelihood of all the responses is accordingly as equation (4).

$$L(\Psi \mid \overline{R}) = \prod_{n=1}^{N} \Phi\left(\delta_n\right)^{R_n} \left(1 - \Phi\left(\delta_n\right)\right)^{1-R_n} \tag{4}$$

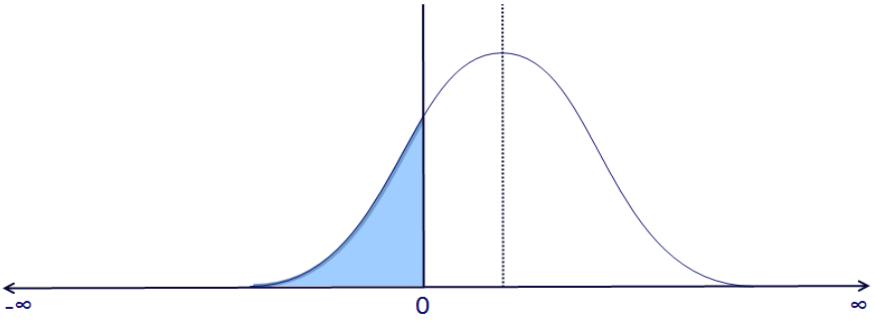

Fig. 4. Probability of second pair being selected

There is no closed form for such a solution, so a direct numerical maximization method needs to be used to compute the estimates (5).

$$\widehat{\Psi} = \arg\max_{\overline{\Psi}} L(\overline{\Psi} \mid \overline{R}) \tag{5}$$

More details on MLDS for video quality can be found in [14] and on MLDS for image quality in [15].

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 their 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 $\overline{\Psi} = f\left(\overline{X}\right)$. 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 $\psi_k - \psi_l \geq \psi_k - \psi_m$ in the psychological domain Figure 5.

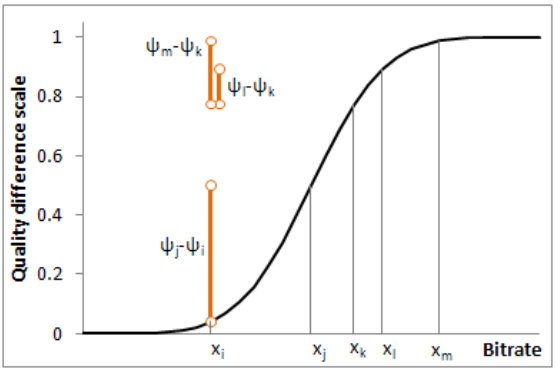

Fig. 5. Monotonicity of the psychometric curve

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 5).

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 $\psi_f - \psi_e > \psi_h - \psi_g$ (Figure 6).

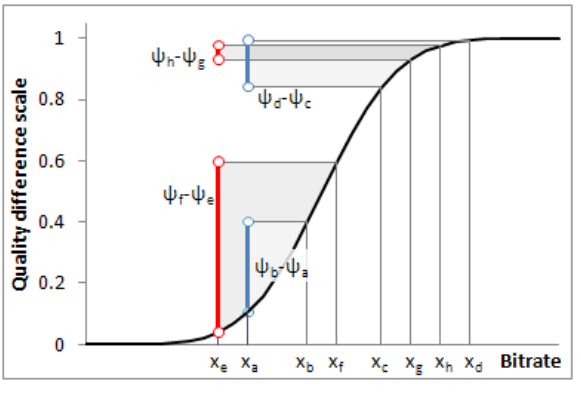

Fig. 6. If first pair in T_1 is bigger than first pair of T_2 is bigger as well

- 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 $\psi_f - \psi_e < \psi_h - \psi_g$.

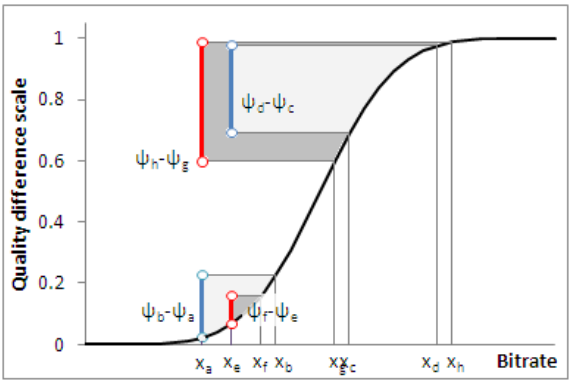

Fig. 7. If second pair in T_1 is bigger than second pair of T_2 is bigger as well

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 given in (6).

$$P(T_1) = \frac{P(T_2)(1-P(T_3))}{P(T_2)(1-P(T_3)) + (1-P(T_2))P(T_3)} \tag{6}$$

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. The software simulates the learning process of the adaptive MLDS algorithm by sequentially introducing data from a previously [14] executed subjective study. The simulation test-bed is a Java application that loads the subjective data from a file, and then sequentially introduces new datapoints. The datapoints are selected by the adaptive MLDS algorithm and the estimated values are used to calculate the psychometric curve in each iteration. The output is compared to the output of running MLDS on the full dataset and 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 [12] is used directly from R using a Java to R interface. To account for the variation in the results due to the random start and random data introduction in the comparison process, the simulation is repeated 100 times and results averaged. Finally the simulation process was computationally very demanding. Each numerical optimization was bootstrapped 1000 times. This was repeated for each step in the introduction of new batch of data and for each video. All this for a single simulation. To handle the computational demand the simulation was executed on a high performance computing grid.

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, which provides for early stopping of the experiment. The performance of the adaptive MLDS is presented in Figure 8, 9, 10 and 11. In Figure 8 we present the accuracy of the estimations for ten types of videos against the number of introduced datapoints. In Figure 9 we observe the leaning rate of adaptive MLDS against the classical 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.

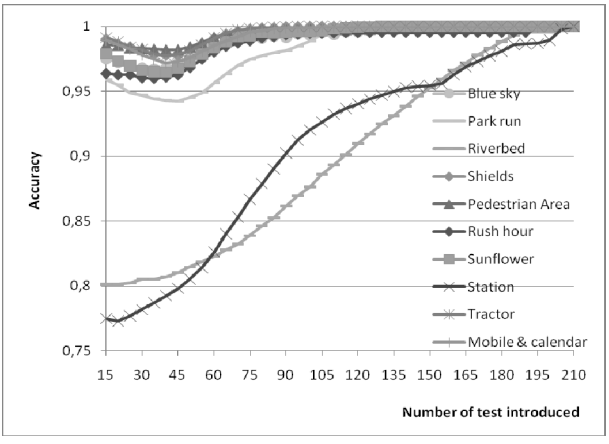

Fig. 8. Accuracy of the estimations

In Figure 10 we present the standard deviation of the different value for the RMSE at each point. Figure 11 presents the distribution of the confidence or the probabilities of those estimations. Starting from the initial 15 data points most of the unknown 195 test are estimated with 0.5 accuracy, but soon after introducing more data the estimations rapidly improve. 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 9. The accuracy of the predicted psychometric curves is high for all

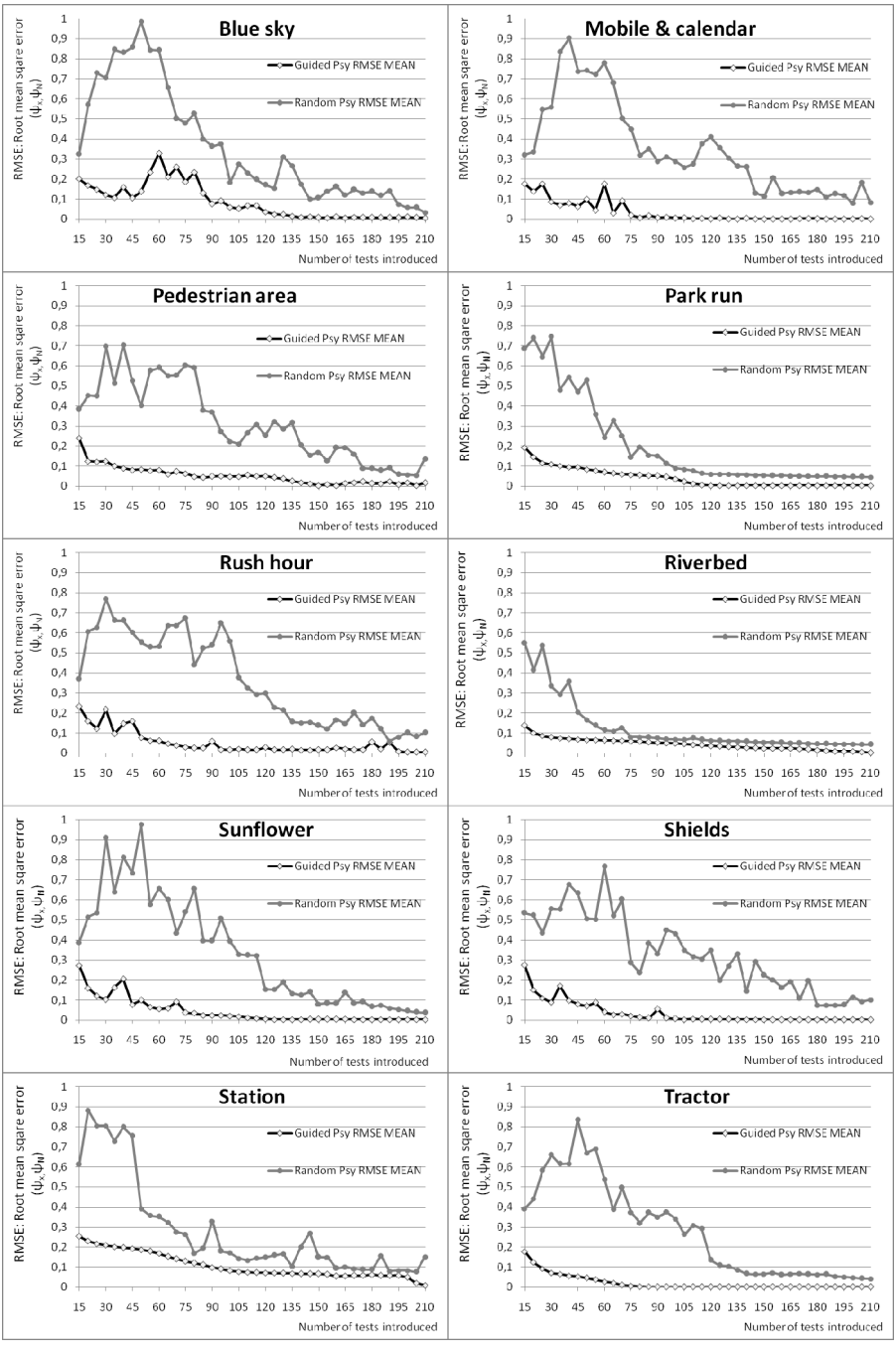

Fig. 9. Mean RMSE for the ten types of video

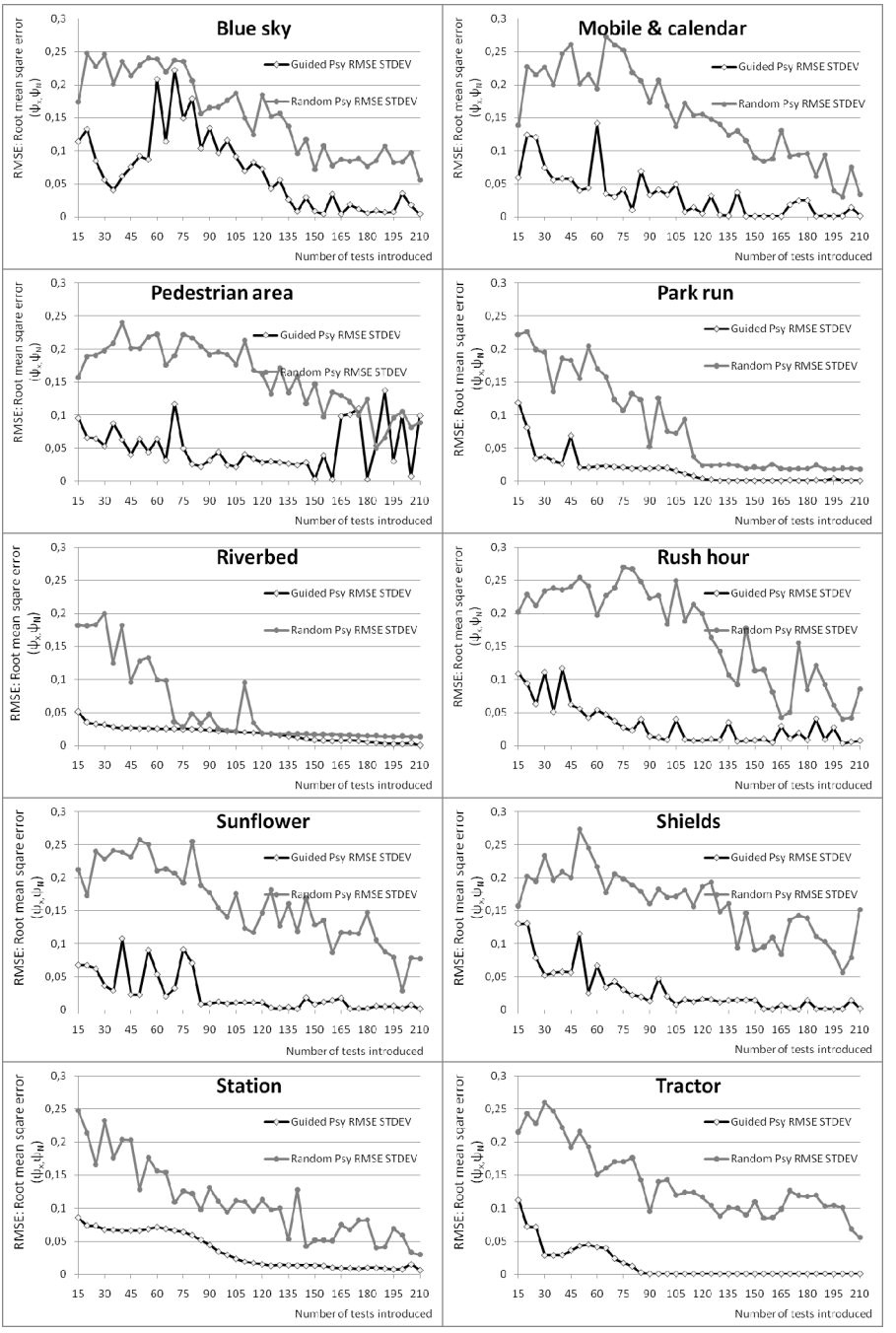

Fig. 10. Standard deviation of the RMSE for the three types of video

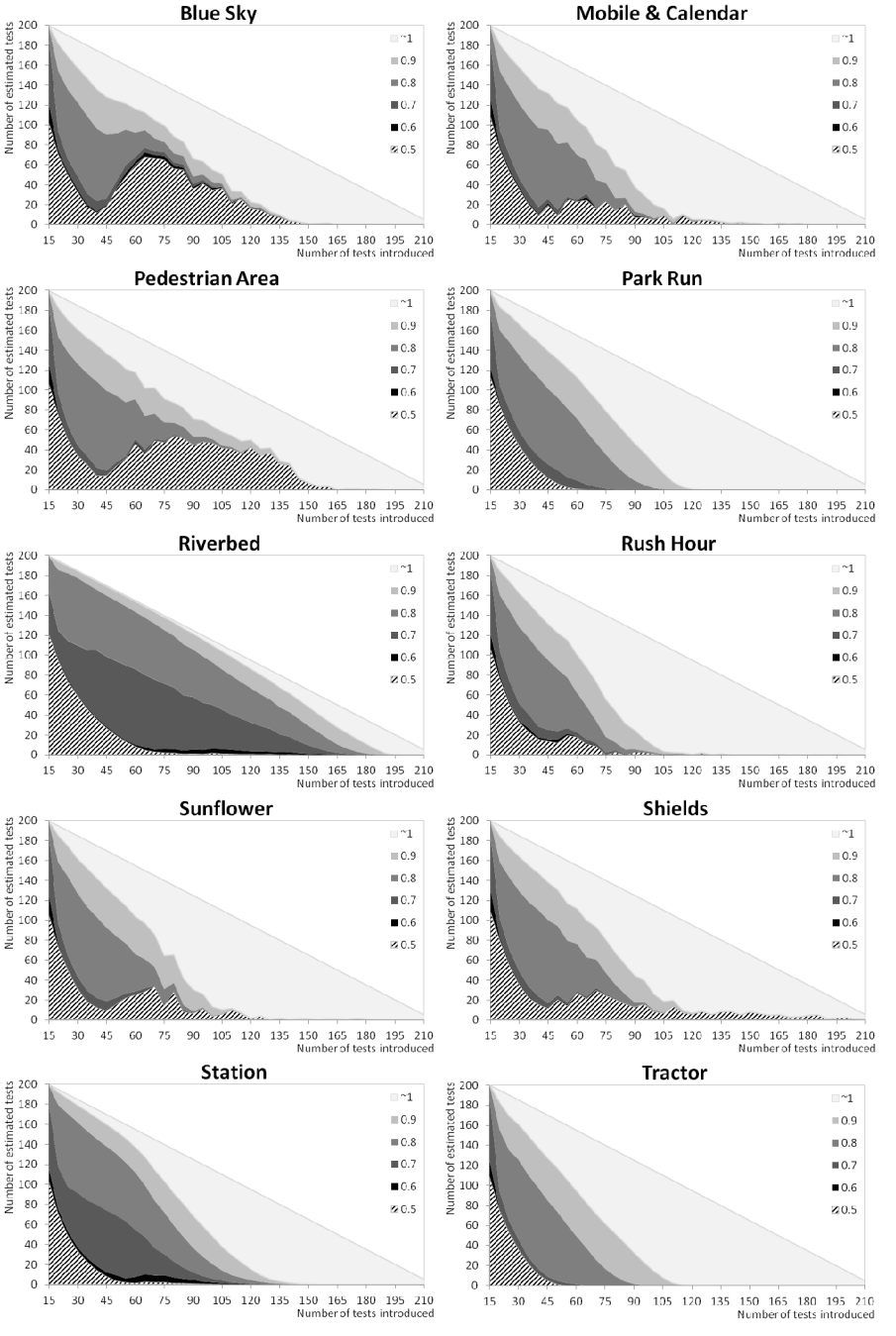

Fig. 11. Estimation confidences for the three types of videos over the number of introduced datapoints

datasets in this range. The RMSE is bellow 0.3 between the 10 given and predicted Ψ values. The accuracy in the prediction is generally very high and improves with the introduction of more data, shown in Figure 8. The Riverbed and Station are more difficult to learn due to high noise in the answers, which makes them also more difficult to estimate.

6 Conclusions

The adaptive MLDS algorithm is an active learning algorithm specifically designed for the MLDS method, a 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 its 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. This further makes this method an excellent candidate for use in management of video delivery services and optimizing the QoE.

Further we intend to apply this method in a subjective study of a more diverse environment, involving different devices and modes of use. Finally the method could be extended in the online learning direction for highly dynamic environments where model validity is short termed.

References

1. Watson, A.B.: Proposal: Measurement of a JND scale for video quality. IEEE G-2.1. 6 Subcommittee on Video Compression Measurements (2000)
2. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: Thirty-Seventh Asilomar Conference on Record of the Signals, Systems and Computers, vol. 2, pp. 1398–1402 (2003)
3. Hekstra, A.P., et al.: PVQM-A perceptual video quality measure. Signal Processing: Image Communication 17(10), 781–798 (2002)
4. Seshadrinathan, K., Bovik, A.C.: Motion-based perceptual quality assessment of video. In: Proc. SPIE-Human Vision and Electronic Imaging (2009)
5. Gunawan, I.P., Ghanbari, M.: Efficient reduced-reference video quality meter. IEEE Transactions on Broadcasting 54(3), 669–679 (2008)
6. Winkler, S., Mohandas, P.: The Evolution of Video Quality Measurement: From PSNR to Hybrid Metrics. IEEE Transactions on Broadcasting 54(3), 660–668 (2008)
7. Kalpana Seshadrinathan, A., Rajiv Soundararajan, B., Alan, C.B.B., Lawrence, K.C.B.: A Subjective Study to Evaluate Video Quality Assessment Algorithms
8. Krantz, D.H., Luce, R.D., Suppes, P., Tversky, A.: Foundations of measurement, vol. 1: Additive and polynomial representations. Academic, New York (1971)
9. Shepard, R.N.: On the status of'direct'psychophysical measurement. Minnesota Studies in the Philosophy of Science 9, 441–490

10. Shepard, R.N.: Psychological relations and psychophysical scales: on the status of direct. Journal of Mathematical Psychology 24(1), 21–57 (1981)
11. Ehrenstein, W.H., Ehrenstein, A.: Psychophysical methods. Modern Techniques in Neuroscience Research, 1211–1241 (1999)
12. Knoblauch, K., Maloney, L.T.: MLDS: Maximum likelihood difference scaling in R. Journal of Statistical Software 25(2), 1–26 (2008)
13. Green, D.M., Swets, J.A.: Signal detection theory and psychophysics (1966)
14. Menkovski, V., Exarchakos, G., Liotta, A.: The value of relative quality in video delivery. Eindhoven University of Technology, Eindhoven (2011)
15. Charrier, C., Maloney, L.T., Cherifi, H., Knoblauch, K.: Maximum likelihood difference scaling of image quality in compression-degraded images. Journal of the Optical Society of America A 24(11), 3418–3426 (2007)