Journal of Mobile Multimedia, Vol.7, No.3 (2011) 151-162

THE VALUE OF RELATIVE QUALITY IN VIDEO DELIVERY

VLADO MENKOVSKI

Eindhoven University of Technology, Eindhoven, The Netherlands
V.Menkovski@tue.nl

GEORGIOS EXARCHAKOS

Eindhoven University of Technology, Eindhoven, The Netherlands
G.Exarchakos@tue.nl

ANTONIO LIOTTA

Eindhoven University of Technology, Eindhoven, The Netherlands
A.Liotta@tue.nl

Received February 16, 2011
Revised July 14, 2011

Estimating perceived quality of video is typically done by gauging the user's response on an absolute scale of ratings (excellent, good, fair, poor and bad). However, the internal representation of these adjectives to the stimuli varies significantly in different people. Even though the goal is to make an absolute estimate of the perceived quality, these questions reveal merely relative tendencies due the incorporated bias and variability in the responses. We present results from quality assessment based on estimates of relative quality distances between samples, by asking the question in the form or comparison rather than rating. This, two-alternative forced choice method scales the differences in a form of psychometric function, which presents the utility of the perceived quality on a measurable objective value. We argue that this relativistic mapping with low variance is more useful in video delivery because it offers an accurate way to optimize the resources.

Key words: Maximum Likelihood Difference Scaling, MLDS, Video Quality, Quality of Experience, QoE
Communicated by: D. Taniar & E. Pardede

1 Introduction

Delivering the desired quality of multimedia content necessitates the understanding of how video quality is perceived by the viewers. Commonly used subjective methods are based on a rating procedure that estimates the video quality on an absolute scale from excellent to poor [1]. However, people's internal representation of such scales is intrinsically biased and varies from person to person. This bias and variance is propagated to the test output and results in the inefficiency of this type of subjective studies. This is not surprising, in fact psychophysicists have argued for a long time that the brain perceptual system is more accurate at grasping 'differences' rather than giving absolute rating values [2].

In this paper we present a method based on difference scaling rather than rating used to quantify the degradation of quality in video as a function of the encoding bit-rate. The method uses MLDS (maximum likelihood difference scaling) [3], a two-alternative-forced-choice (2AFC) procedure from the domain of psychophysics. This procedure utilizes the mechanisms of direct comparison rather than rating that are significantly less biased and with low variance. Because of their characteristics the 2AFC methods offer more accuracy with less testing [2].

To explain the mechanics of difference scaling we present an example of our MLDS analysis in Figure 1. We analyzed a video for degradation of quality due to H.264 encoding with constant bit-rate. The bit-rate values range from 2Mbps to 64kbps. The figure represents the normalized relative difference between the 2Mbps video and the rest of the videos. The higher the value, the bigger the difference in quality. The MLDS does not provide direct estimates of quality such as ‘good’ or ‘poor’, rather it only provides for relative differences between the samples. However, these relative differences offers significant understanding of how the impairment affects the perceived quality. For example in the figure it is evident that the relative distance between the quality of the 2Mbps sample and the 512kbps video is almost zero. This leads to the conclusion that the benefit of increasing the bit-rate from 512kbps to 2Mps is close to none. Further we can make observations as: the distance between the 256kbps and 128kbps is almost the same as the 128kbps and 64kbps. This means that the gain of increasing the bit-rate from 64kbps to 128kbps is the same as increasing it from 128kbps to 256kbps. In turn, we can estimate the utility of the first increase would be double than the second, assuming that the price per bit is constant.

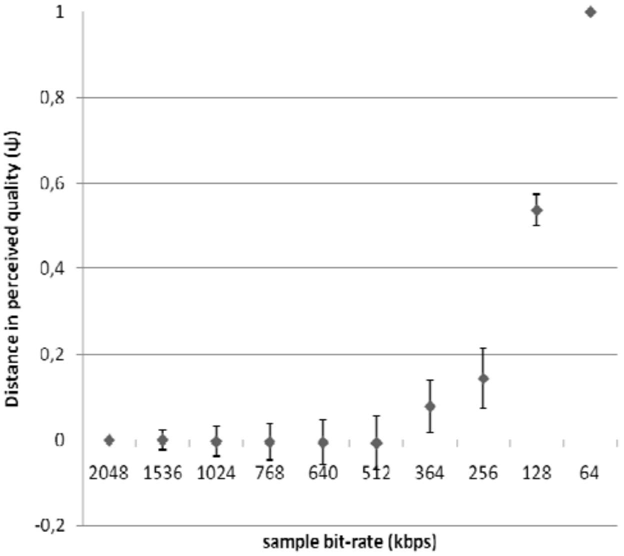

Figure 1. Differences in quality MLDS results

Understanding how a resource affects the perceived quality allows for estimating the utility of that resource. In the given example, the utility halves from the first to the second increment, and fully diminish above 512kbps. Having a grasp of the resource utility in video delivery allows for optimizing the service by providing ‘constant quality’ instead of ‘constant resources’ to the viewers. This is achievable without the need to exactly know if the video quality was perceived as ‘good’ or ‘poor’, but only by understanding the relative distances in quality.

In this paper we present the details of the MLDS method and we demonstrate its applicability and advantages by executing a subjective video quality study and elaborating on the results.

2 Background

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 [4]. 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. [5] 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. In the subjective study they implemented complex procedures to deal with contextual and memory affects as well as unreliable (biased) subjects.

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 [6][7][8][9][10].

Research done is psychophysics, a discipline that quantifies the effect on stimuli on internal perception, has established the m-AFC testing as a primary estimation procedure [11] for quantifying intensity of stimuli. The 2AFC methods present the person with a choice of two stimuli and ask him to discriminate between the intensity of the two. These types of tests have less bias and variability because the procedure is more natural and direct to a person; no internal mapping is necessary.

The MLDS method in the literature has been utilized for estimation of quality differences for images. Charrier et al. in [12] present quality difference scaling of compressed images with a lossy image compression techniques. They implement a comparison of image compression in two different color spaces, and conclude that in the CIE 1976 L*a*b* color space the images can be compressed by 32% more, without additional loss in quality. Their results and discussion clearly show the applicability of MLDS and the ease of collecting data with it.

In this paper we use MLDS to estimate the quality scale for a range of videos with different spatial and temporal characteristics. The results presented demonstrate that MLDS can be used for estimating quality of video with higher accuracy and significantly lower testing costs than subjective rating.

3 Maximum likelihood difference scaling

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 [11] as depicted in Figure 2.

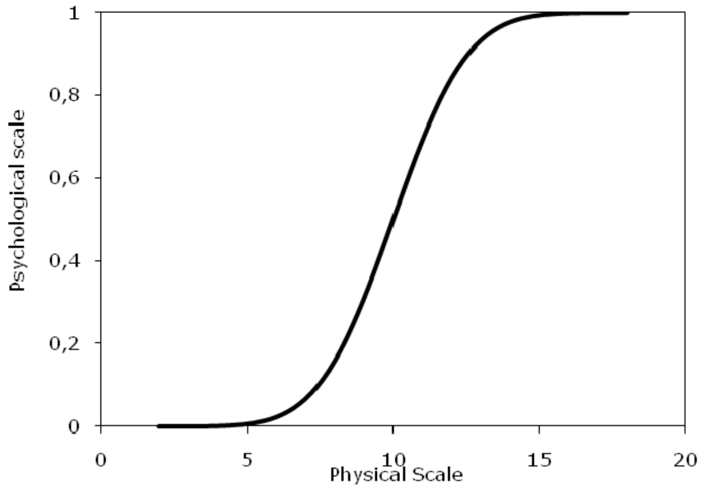

Figure 2. Psychometric function

The horizontal axis of the Figure 2 represents the physical intensity of the stimuli – in our study this will be the video bit-rate. The vertical axis represents the psychological scale of perceived difference in signal strength – for our purpose the difference in video quality. The perceptual intensity of the first (or reference) sample ψ_1 is 0 and the last sample perceptual difference ψ_{10} is fixed to 1 without the loss in generality [13]. The MLDS produced model is an estimate of the rest of the parameters of the viewers' internal quality scale.

The 2AFC test is designed in the following manner. Two pairs of videos are presented to the viewers (ψ_i, ψ_j and ψ_k, ψ_l). The intensity of the physical stimuli is always in the following manner $i < j$ and $k < l$. The method needs to compare sizes of distances between the qualities of videos so that the results can directly let us build a model of the quality distance between all of the presented videos.

The viewer needs to select the pair of videos that have bigger difference in quality between them. In other words if the expression $|\psi_j - \psi_i| - |\psi_l - \psi_k| > 0$ is true the viewer selects the first pair, otherwise he or she will choose the second.

Because the stimuli are ordered as $i < j$ and $k < l$ we can assume that in the psychological domain also $\psi_j \geq \psi_i$ and $\psi_l \geq \psi_k$ and we drop the absolute values.

The decision variable used by the observer is the following:

$$\Delta(i,j,k,l) = \psi_j - \psi_i - \psi_l + \psi_k + \varepsilon \tag{1}$$

where ε is the error or noise produced by the viewers visual and cognitive processing. As defined in (1) the observer will select the first pair when $\Delta(i,j,k,l)>0$ or the second when $\Delta(i,j,k,l)<0$.

In order to use the maximum likelihood method to determine the $\Psi=(\psi_1,\ldots,\psi_{10})$ parameters we need to define the likelihood (probability given the parameters) that the viewer will find the first pair with larger difference than the second pair. For this the method models the perceived distances using signal detection theory (SDT) [14].

The equal variance Gaussian model from the SDT is used to model the process of selection that the user is executing for each presented pair. This model assumes that the signal is contaminated with ε, a Gaussian noise with zero mean and standard deviation of σ (Figure 3). Each time the observer is presented with a pairs of videos, the perceived difference is a value of the random variable X drawn from the distribution given in Figure 3. The distribution in Figure 3 is with arbitrary signal strength of 1.

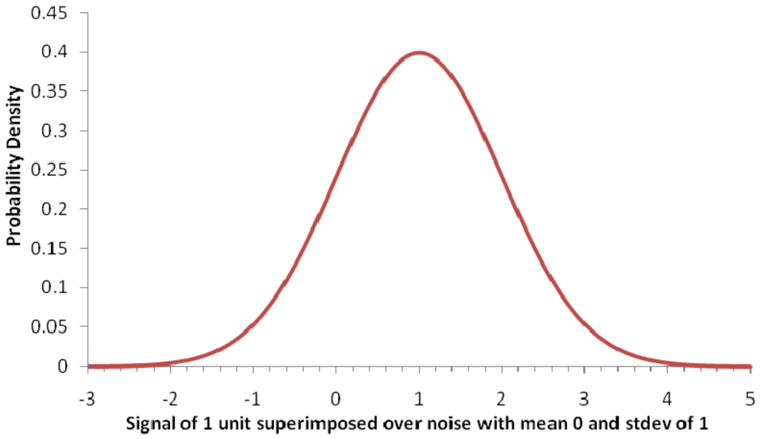

Figure 3. Signal of 1 unit superimposed over noise with 0 mean and standard deviation of 1

The probability that Δ(*i,j,k,l*)>0 is given by the surface under the Gaussian from zero to plus infinity (Figure 4).

For reasons of mathematical simplicity it is better to represent the surface under the curve with a cumulative Gaussian function. The inverse portion of the surface (Figure 5) is as (2).

$$F(x;\mu,\sigma^2)=\Phi\left(\frac{x-\mu}{\sigma}\right)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-(t-\mu)^2/2}\,dt \tag{2}$$

Looking at the inverse part of the surface under the Gaussian the probability of detecting the signal would be:

$P(R=1;\mu_s,\sigma^2)=1-\Phi\left(\frac{0-\mu_s}{\sigma}\right)=\Phi\left(\frac{\mu_s}{\sigma}\right)$ and $P(R=0;\mu_s,\sigma^2)=1-P(R=1;\mu_s,\sigma^2)=1-\Phi\left(\frac{\mu_s}{\sigma}\right)$

where μ_s is the mean or the intensity of the signal, σ is the standard deviation of the noise and R is 1 when the first pair is selected and 0 when the second pair is selected.

The likelihood for the whole set of responses in a test is the product of all of the individual probabilities $L(\Psi,\sigma)=\prod_{n=1}^{N}\Phi\left(\frac{\delta(i,j,k,l)_n}{\sigma}\right)^{R_n}\left(1-\Phi\left(\frac{\delta(i,j,k,l)_n}{\sigma}\right)\right)^{1-R_n}$ where $\delta(i,j,k,l)_n=\psi_{j_n}-\psi_{i_n}-\psi_{l_n}+\psi_{k_n}$.

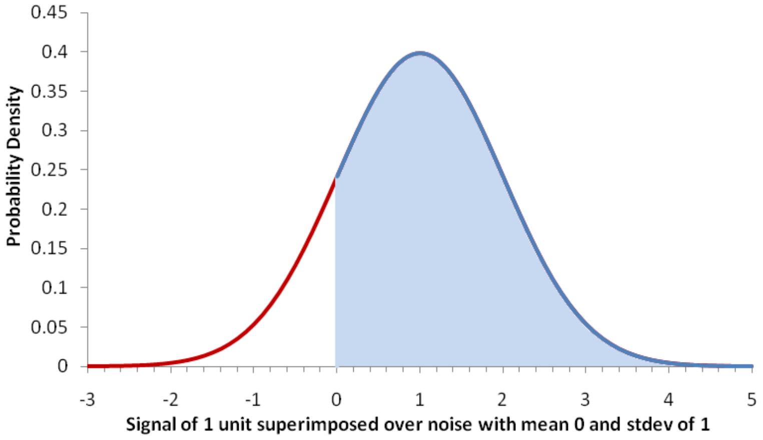

Figure 4. Probability that the signal is positive

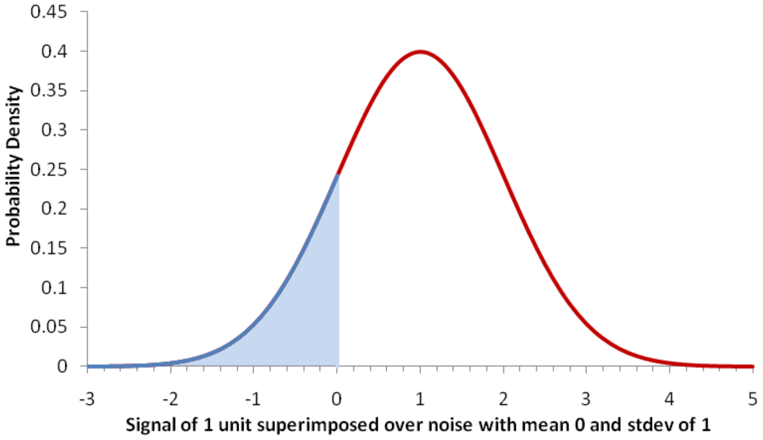

Figure 5. Probability that the signal is negative

The Maximum Likelihood method estimates the parameters, such that the given likelihood is maximized.

For example, if we have $x = \{x^t\}(t = 1..N)$ instances drawn from some probability density family $p(\mathrm{x} \mid \theta)$ defined up to parameters θ (3).

$$x^t \sim p(x \mid \theta) \tag{3}$$

If the x^t samples are independent, the likelihood parameter θ given a sample set x is the product of the likelihood of individual points (4).

$$L(\theta \mid x) \equiv p(x \mid \theta) = \prod_{t=1}^{N} p(x^t \mid \theta) \tag{4}$$

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

$$\hat{\theta} = \text{argmax}_{\theta}\, l(\theta \mid x) \tag{5}$$

4 Experimental setup

The experimental setup consists of a web application that displays the two pairs of videos to the viewer in the layout given in Figure 6. The user response is recorded in the application database. The web application is developed using the java server pages technology [15]. The videos are displayed using the JW player [16], which is a Flash 5 player that is capable of displaying H.264 encoded videos.

The videos are encoded using the X264 library **[17]** and saved in mp4 file format. The raw videos are the unimpaired samples of the Live video database **[5][18]** used for subjective studies of video quality.

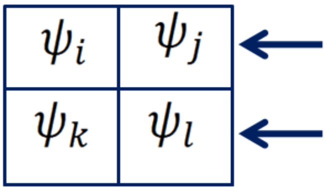

Figure 6. Four video displayed

The ten different videos (Table 2, Figure 7) are encoded with constant bit-rate of ten different values ranging from 2Mbps to 64kbps. The videos have 25 frames per second frame-rate and a spatial resolution of 768 by 432 pixels

The video player is configured to pre-buffer the full content before playing, so additional impairments such as freezes during the playback are avoided.

Table 2. List of the video descriptions

bs	Blue Sky	Circular camera motion showing a blue sky and some trees
rb	River Bed	Still camera, shows a river bed containing some pebbles in the water
pa	Pedestrian Area	Still camera, shows some people walking about in a street intersection
tr	Tractor	Camera pan, shows a tractor moving across some fields
sf	Sunflower	Still camera, shows a bee moving over a sun-flower in close-up
rh	Rush hour	Still camera, shows rush hour traffic on a street
st	Station	Still camera, shows railway track, a train and some people walking across the track
sh	Shields	Camera pans at first, then becomes still and zooms in; shows a person walking across a display pointing at it
mc	Mobile & Calendar	Camera pan, tor train moving horizontally with a calendar moving vertically in the background
pr	Park run	Camera pan, a person running across a park

The results are collected in a database in the format:

bit-rate 1	bit-rate 2	bit-rate 3	bit-rate 4	R (index bigger pair)

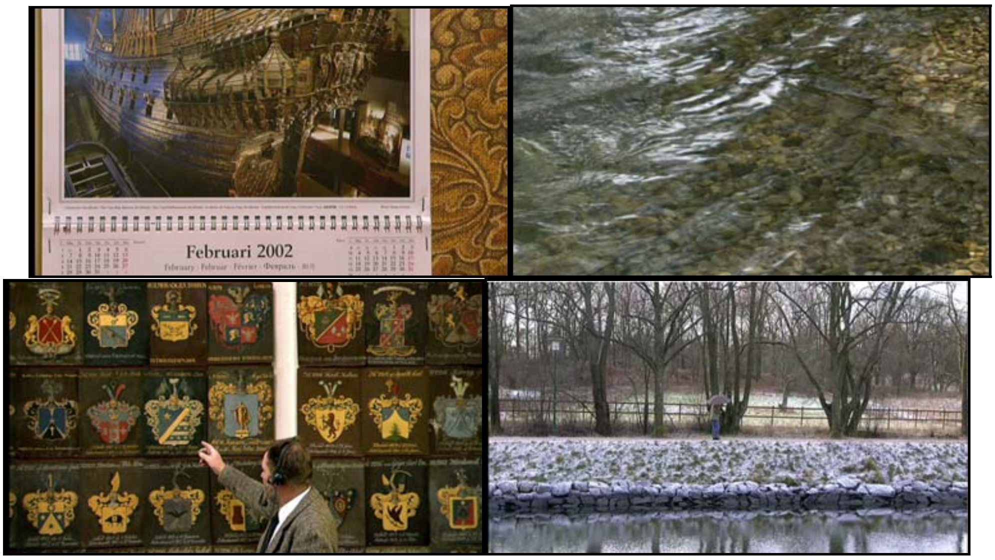

Figure 7. Snapshots of the mc, rb, sh and pr video

We used the MLDS implementation **[13]** in R **[19]** to calculate the $\Psi=(\psi_1,...,\psi_{10})$ values. The output ψ values are fitted to a psychometric curve using a probit regression fit with variable upper/lower asymptotes using the 'psyphy' package in R **[20]**.

The results of the subjective study are μ and σ of a cumulative Gaussian curve for each type of video that models the relationship between the video bit-rate and perceived quality for each type of video**.**

5 Results

The MLDS experiment with 10 levels of stimuli requires 210 responses to cover all possible combinations for a single video. We have done 3 rounds per video sample or 630 tests for each video; in total we have collected 6300 responses. The videos are displayed one at the time or in pairs. They are 10 seconds long, so to view a single test up to 40 seconds are needed, but in most cases the larger difference is evident much sooner to most observers.

To calculate the standard error we executed a bootstrap [21] fitting procedure with 10,000 rounds. The mean values and the standard error are given in Figure 8 and the standard deviation for each point in Figure 9.

The results in Figure 8 show that most of the videos follow a similar trajectory of the difference in quality. There is little perceived difference down to 512kbps and then a rapid rise appears. The difference is not zero in the high range, as we can also see from the standard error on the points from 1536kbps to 512kbps, but it is very low relative to the lower bit-rate samples. This means it is safe to say that there is little benefit from increasing the bit-rate above 512kbps. The exception is the 'rb' video and somewhat the 'pr' video. The 'rb' video displays a surface of water, which shows significantly different compression characteristics than the rest of the videos.

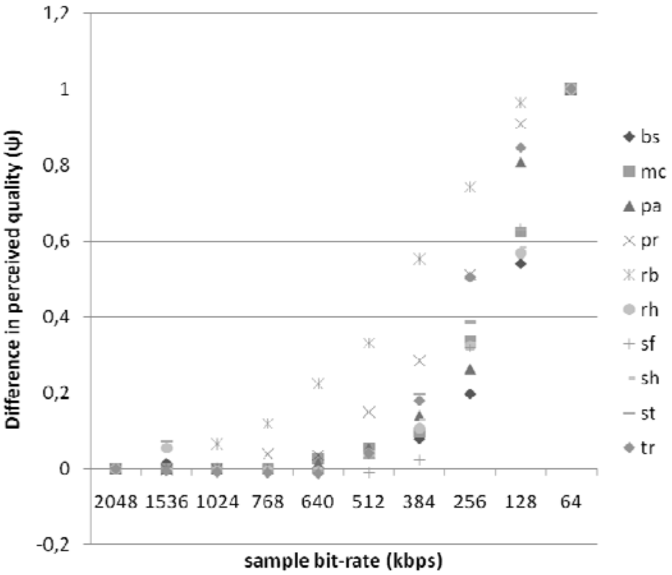

Fig. 8. Results of the MLDS experiment by video type

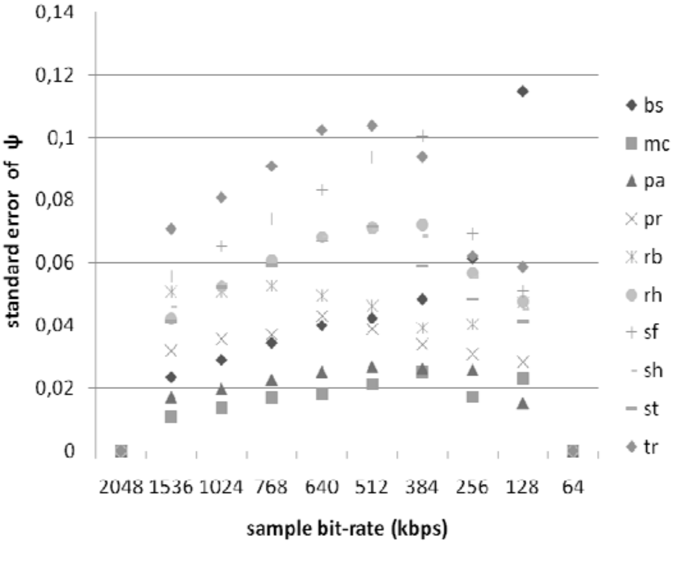

Fig. 9. Standard error of the MLDS results by video type

To quantitatively analyze the characteristics of each model we fitted a cumulative Gaussian curve to each difference model as demonstrated in Figure 10, which represents the psychometric curve [22]. There is high goodness of fit to the curve with small residuals. This further demonstrates the success of this subjective study to model the quality difference perception with a psychometric curve.

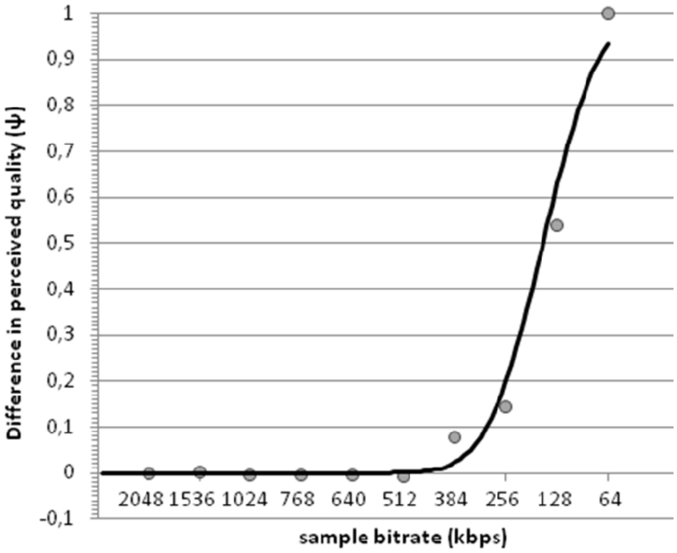

Figure 10. Fitting a cumulative Gaussian on the bs MLDS model

For each video the μ and σ of the fitted curve are given in Table 2. A plot of each of the fitted models is given in Figure 11. The plotted curves model a smooth quality distance for different bit-rates from the reference 2Mbps video.

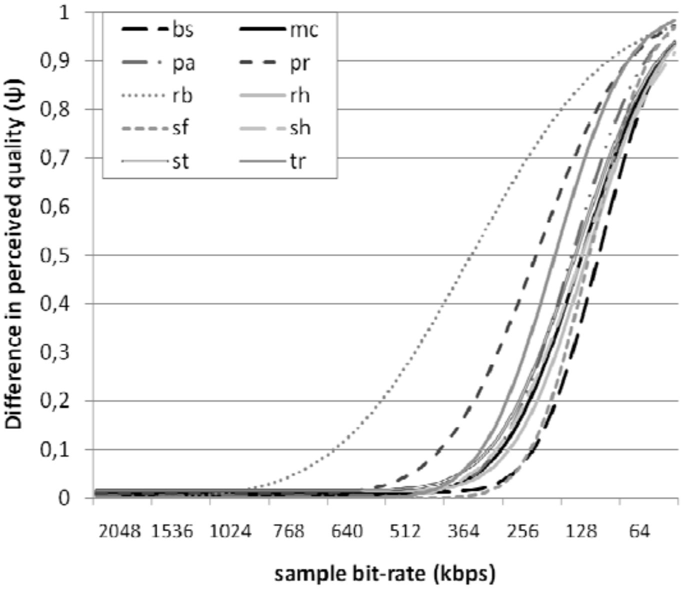

Fig. 11. Fitted psychometric curves from the MLDS results

Table 2. The μ and σ of the cumulative Gaussian

bs	mc	pa	pr	rb	rh	sf	sh	st	tr
-5.43	-5.07	-5.08	-4.57	-4.09	-5.22	-5.54	-5.13	-5.00	-4.94
0.24	0.20	0.20	0.15	0.11	0.22	0.25	0.21	0.20	0.19

Observing the parameter values in Table 2 we can make the same conclusions from above in a quantitative form. The mean value of the psychometric curve of the ‘rb’ video is noticeably lower than the rest of the videos, so the distance increases earlier than in with others. The remaining psychometric curves cluster together and confirm that most of these videos difference in quality is negligible to the reference down to 512kbps, while the bit-rate between 256 and 128kbps is half way to the perceived distance between the reference and the 64kbps video. The results accurately capture the nonlinearity in the perceived quality by the viewers.

6 Conclusions

In a subjective study we want to estimate how people perceive the quality exactly rather than scaling the differences, which makes methods like MLDS seem counter intuitive. However, we have shown that relative scaling of the difference can be just as much useful because it produces a utility function for the resource that the quality depends on. In addition, MLDS does not suffer from the common pitfalls associated with rating procedures, mostly due to the fact that comparison comes more naturally to people rather than rating. Consequently, the training of the participants for a MLDS test is simple and straight-forward. Moreover, for a significant number of tests the difference in quality is obvious and the time to collect the answer is small. Simpler training and shorter tests bring the amount of time and costs for the subjective study down. The high confidence in many of the participant responses allows for lower standard error and variation over the results and improvement in the goodness of fit for the models.

The method can be also applied to measuring degradation of quality due to other factors like transport impairments i.e. bit-error rate or IP packet loss.

In continuation of this work, we plan to use this method in subjective estimation of different compression algorithms. Even though a direct comparison is not possible, using MLDS, the dynamic range of compression over bit-rate can be examined and compared.

References

1. R. I. T. U. R. P. ITU-T, “910,” *Subjective video quality assessment methods for multimedia applications*, 1999.
2. A. B. Watson, “Proposal: Measurement of a JND scale for video quality,” *IEEE G-2.1. 6 Subcommittee on Video Compression Measurements*, 2000.
3. L. T. Maloney and J. N. Yang, “Maximum likelihood difference scaling,” *Journal of Vision*, vol. 3, no. 8, 2003.
4. 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.
5. A. Kalpana Seshadrinathan, B. Rajiv Soundararajan, C. B. B. Alan, and K. C. B. Lawrence, “A Subjective Study to Evaluate Video Quality Assessment Algorithms.”
6. D. H. Krantz, “A theory of context effects based on cross-context matching* 1,” *Journal of Mathematical Psychology*, vol. 5, no. 1, p. 1–48, 1968.

7. D. H. Krantz, “A theory of magnitude estimation and cross-modality matching* 1,” *Journal of Mathematical Psychology*, vol. 9, no. 2, p. 168–199, 1972.
8. D. H. Krantz, R. D. Luce, P. Suppes, and A. Tversky, “Foundations of measurement, vol. 1: Additive and polynomial representations,” *New York: Academic*, 1971.
9. R. N. Shepard, “On the status of’direct'psychophysical measurement,” *Minnesota studies in the philosophy of science*, vol. 9, p. 441–490.
10. R. N. Shepard, “Psychological relations and psychophysical scales: On the status of,” *Journal of Mathematical Psychology*, vol. 24, no. 1, p. 21–57, 1981.
11. W. H. Ehrenstein and A. Ehrenstein, “Psychophysical methods,” *Modern techniques in neuroscience research*, p. 1211–1241, 1999.
12. 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.
13. 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.
14. D. M. Green and J. A. Swets, “Signal detection theory and psychophysics,” 1966.
15. H. Bergsten, *JavaServer pages*. O’Reilly & Associates, Inc. Sebastopol, CA, USA, 2003.
16. “JW Player.” [Online]. Available: http://www.longtailvideo.com/players/jw-flv-player/. [Accessed: 29-Nov-2010].
17. L. Aimar et al., “x264-a free h264/AVC encoder,” *Online (last accessed on: 04/01/07): http://www. videolan. org/developers/x264. html.*
18. K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, “Study of subjective and objective quality assessment of video,” *Image Processing, IEEE Transactions on*, vol. 19, no. 6, p. 1427–1441, 2010.
19. R. Team, “R: A language and environment for statistical computing,” *R Foundation for Statistical Computing Vienna Austria ISBN*, vol. 3, no. 10, 2008.
20. K. Knoblauch, M. K. Knoblauch, and M. Suggests, “Package ‘psyphy’.”
21. B. Efron and R. J. Tibshirani, “An Introduction to the Bootstrap: Monographs on Statistics and Applied Probability, Vol. 57,” *Chapmann and Hall, New York*, 1998.
22. F. A. Wichmann and N. J. Hill, “The psychometric function: I. Fitting, sampling, and goodness of fit,” *Perception & Psychophysics*, vol. 63, no. 8, p. 1293, 2001.