

Detection of Distortion in Small Moving Images, Compared to the Predictions of a Spatio-Temporal Model

Kjell Brunnström^{*}, Bo N. Schenkman[†], Albert J. Ahumada Jr.[‡]

^aACREO AB, Electrum 236, SE-164 40 Kista, Sweden

^bNASA Ames Research Center, Moffett Field, CA 94035-1000

ABSTRACT

The image sequence discrimination model we use models optical blurring and retinal light adaptation. Two parallel channels, sustained and transient, with different masking rules based on contrast gain control, are used. Performance of the model was studied for two tasks representative of a video communication system with versions of monochrome H.263 compressed images[§]. In the first study, five image sequences constituted pairs of non-compressed and compressed images to be discriminated with a 2-alternative-forced-choice method together with a staircase procedure. The thresholds for each subject were calculated. Analysis of variance showed that the differences between the pictures were significant. The model threshold was close to the average of the subjects for each picture, and the model thus predicted these results quite well. In the second study, the effect of transmission errors on the Internet, i.e. packet losses, was tested with the method of constant stimuli. Both reference and comparison image was distorted. The task of the subjects was to judge whether the presented video quality was worse than the initially seen reference video. Two different quality levels of the compressed sequences were simulated. The differences in the thresholds among the different video scenes were to some extent predicted by the model. Category scales indicate that detection of distortions and overall quality judgements are based on different psychological processes.

Keywords: video, image quality, spatio-temporal, vision model, H263, packet loss, Internet

1 INTRODUCTION

The Internet provides a huge infrastructure for connecting people in inexpensive ways over large distances. Services such as telephony and video conferences, are becoming available to the ordinary customer. However, the quality is still poor, especially image[§] quality for video conferences. This is due to bandwidth limitations and packet-based transmission. Bandwidth limitations will force high levels of compression, and packet-based transmission can reduce control over the packet arrival time. In addition, packets may be lost due to network congestion. Delayed packets can be included or discarded upon arrival, but in either case, they introduce errors at the receiving end. Standards for giving priority to certain packets are under development and this will certainly decrease the delays and losses. However, there will most likely be a cost for using this type of transmission. The customer may then be provided with a quality level that they can afford. One approach to ensuring good or at least satisfactory image quality is to use a visual model to compare reference images of acceptable quality with the transmitted images. In this study we measure the detection by viewers of poorer quality and see whether this detection can be predicted by a visual model.

There have been many reports at earlier Electronic Imaging conferences of similar efforts to model the early-vision system and use the model in technical applications. Examples of models aimed at video applications are those presented by Watson et al. (1999)¹ and by Winkler (1999)². The present article describes the success of such a model in predicting the detection of image compression distortion in image sequences. We use the spatio-temporal visual model that was presented earlier by Ahumada et al. (1998)³, who evaluating its performance for contrast sensitivity and masking. Another study compared the predictions of the model with human performance of target detection in moving infrared images (Brunnström et al. 1999)⁴. One of our intentions in the present experiments was to test this model for video applications. This image sequence discrimination model has processing stages representing optical blurring and retinal light adaptation. The processing then

^{*} Correspondence: Email: kjell.brunnstrom@acreo.se, Tel: +46-8-6327732, Fax: +46-8-7505430, URL: www.acreo.se

[†] Email: bo.schenkman@acreo.se

[‡] Email: al@vision.arc.nasa.gov, URL: vision.arc.nasa.gov/~al/ahumada.html

[§] We will use the word 'image' to denote image sequence, moving image or video, while non-moving images will be denoted by e.g. 'still image' or 'individual frame'.

proceeds in two parallel channels, one called Magno, responding to higher temporal and lower spatial frequencies (the transient channel) and one called Parvo, more sensitive to low temporal and high spatial frequencies (the sustained channel). This division simulates the separation of processing in the ganglion cells and in the Magno and Parvo structures in the Lateral Geniculate Nucleus. Following these filtering operations, separate and different masking operations, based on contrast gain control, are applied in the two channels.

In this article we describe two experiments designed to assess the utility of the model for predicting detection of increased video communication system errors. In the first experiment the errors were video compression errors, the result of bandwidth limitations.

In the second experiment the transmission errors were the result of lost packets, as might occur on the Internet. In Experiment 1 we used a 2-alternative forced choice method, with the test and reference sequences shown together temporally. This is a common psychophysical method when studying detection. Usually, however, a user does not have access to a reference image, does not know of the image quality of the original, or may have seen it some time ago, and makes a comparison with a remembered image. The method of Experiment 2 incorporates this memory aspect. The viewer here had to compare the image presented with a remembered image. One goal of Experiment 2 was to see if this memory method was a useful method for understanding the image quality problems of transmission systems

How does one evaluate the ability of a model to predict user image quality judgments? One may ask how this model compares to other, maybe simpler physical measures, such as the actual number of packets lost or the simple Peak Signal to Noise Ratio (PSNR). A good model should preferably be substantially better than these simpler measures. The second aim of Experiment 2 was to study the relative different explanatory power of the different physical measures.

The use of category scales in psychophysics was criticized by S.S. Stevens and by G. Ekman (see Borg, 1982)⁵. They only approved of methods resulting in ratio scales. Martens and Boschmann (in press)⁶ advocate the use of category scales as a method of understanding image quality. The use of grading scales is also included in assessment procedures for television pictures (Rec.ITU-R BT.500-7)⁷. In the realm of visual display units, Roufs and Boschman (1997)⁸ found that numerical category scaling offered a fast and efficient method for measuring the psychological attribute “visual comfort”. The last aim of Experiment 2 was to find more global characteristics of the judged image quality by the use of category scales.

2 EXPERIMENT 1: VIDEO TRANSMISSION

2.1 EXPERIMENTAL METHOD

Five image sequences were used to generate pairs to be discriminated. Only the luminance parts were used. In each presented sequence pair, one of the sequences was not compressed, while the other was the same sequence compressed to a varying degree. The compression was made according to the H.263 standard (ITU-T, 1998)⁹. The task of the subject was to identify which of two sequences that was distorted. The psychophysical method was 2-alternative forced choice in combination with a staircase procedure adjusting the distortion level. There were three male subjects with normal vision.

2.2 RESULTS

The just-detectable distortion thresholds were calculated for each subject at each of the image sequences. We also viewed the predictions of the model as a virtual subject, and these values were analyzed together with those for the real persons. Analysis of variance, using the subject by image sequence interaction (12 degrees of freedom) as the error term, showed F-ratios of 9.5 with 4 degrees of freedom in the numerator and 12 degrees of freedom in the denominator, i.e. $F(4,12)=9.5$, for the differences between the pictures and $F(3,12)=7.0$ for the differences between the ‘subjects’. These are both significant at $p=0.05$. The model response was close to the average of the subjects, see Figure 1. When computing a-priori tests of the difference between the means, by t-tests, the threshold of the model did not differ significantly from that of the mean for the three subjects for any of the five image sequences.

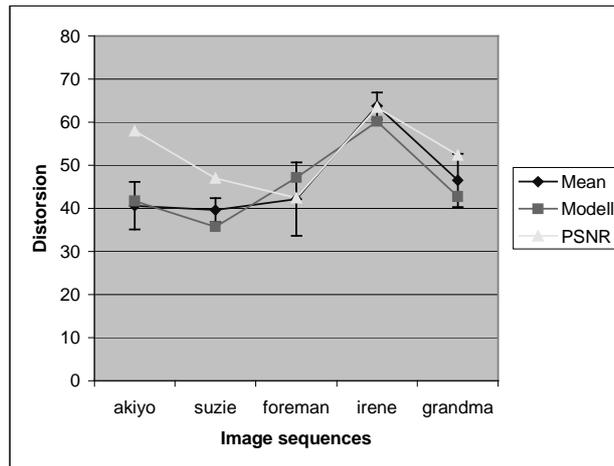


Figure 1: The mean thresholds of the subjects in Experiment 1 and the predicted thresholds of the model. The vertical bars for the means show 95% confidence intervals.

2.3 DISCUSSION

For three images the predictions of the model were close to the mean of the three subjects. For two of the images the fit was less good. The variance of these two images was lower than for the other three, which makes the predictions fall on the border of the 95% confidence interval, although the difference is not greater in absolute values. Although more complex, cognitive mechanisms may be important for some sequences, the model does predict the results quite well. A more detailed description of this experiment may be found in Brunström, Eriksson, Schenkman and Andrén (1999)⁴. The data in Experiment 1 was planned to be a small pilot study and is based on relatively few data. The results and conclusions should be viewed with caution. To investigate the generality of the results, Experiment 2 was conducted with more subjects and a greater number and a greater variety of image sequences.

In real time video transmission today on the Internet, compressed images are transmitted as packets. We were therefore interested in seeing how useful the model would be with this type of image distortion. Furthermore, the model used in this study was constructed to predict the threshold for detection of any difference between two image sequences. For more complex issues, such as more global characteristics of image quality, one may expect that cognitive aspects will be of importance. We intended to measure this aspect of image quality by using category scales. These two questions were addressed in Experiment 2.

3 EXPERIMENT 2: PACKET LOSSES

3.1 SCENES/STIMULI

Images, i.e. scenes were compressed according to the H.263 standard using two layers, one base layer and one enhancement layer. The enhancement layer gives the quality of the images when no packets are lost. Six different scenes were compressed with the variation of the quantization parameters, for the two layers. For one set these were set to 26 and 8, while for the other it was set to 18 and 4. The first value refers to the base layer and the second value to the enhancement layer. These two combinations are here called presentation levels and will subsequently be denoted 18_4 and 26_8.

For each scene the probability of packet loss was varied from 5 to 35% in seven equal steps. For simulating the packet loss it was assumed that the base layer could be transmitted without loss and that the header of the first frame in the sequence is not lost. In order to limit the effects of accidental placements of a certain artifacts in the images, five different versions or instances of each image sequence were generated, for a certain packet loss probability. All images were in black and white.

The reference images, i.e. scenes with packet loss probability of 5% are shown in Figure 2 and 4 for the two presentation levels 18_4 and 26_8, respectively. As a comparison, we have chosen to present for one image, Mother and Daughter, images at the 18_4 presentation for 20 and 35% packet loss, see Figure 3.

Three of the scenes, Akiyo, Mother-and-Daughter and Salesman are so called head and shoulder images and were chosen to represent probable scenes in telecom situations. The other three, i.e. Hall, Jurassic and Stefan were chosen to represent images that could occur in surveillance situation or in entertainment activities, e.g. on the Internet.



Figure 2: The tenth frame of the images, Akiyo, Hall, Jurassic, Mother and Daughter, Salesman and Stefan at the presentation level 18_4 at the packet loss of 5%.

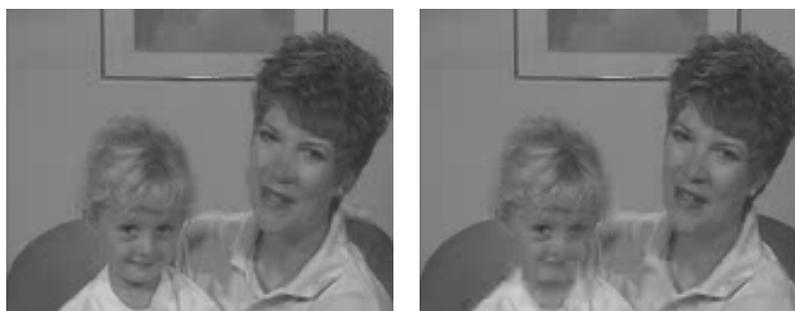


Figure 3: The tenth frame of the Mother and Daughter image at the presentation level 18_4 for the 20% and 35% packet loss, left and right, respectively.

3.2 ROOM CONDITIONS

The participant sat in a small chamber with gray homogeneous cloth surfaces, both in front, above and to the sides of him or her. The room illuminance on the screen was 96 lx, measured in the horizontal plane and centrally on the screen. The outer surface of the monitor and the table in front of the person were also covered with the gray cloth.



Figure 4: The tenth frame of the images, Akiyo, Hall, Jurassic, Mother and Daughter, Salesman and Stefan at the presentation level 26_8 at the packet loss of 5%.

3.3 APPARATUS

The monitor used was Eizo, 17 inch, model T562-T in 800x600 resolution. The maximal luminance level for gray level 255 was set equal to 100 cd/m². The resulting gamma function was measured with this original setting. The picture frequency was 75 Hz. The active area of the screen had a width of 324 mm and a height of 243 mm. The image sequence presented on this screen was horizontally 64 mm (6.5 deg) and 51 mm (5.2 deg) vertically. The participant sat at a distance of 0.56 m from the screen, with his chin on a chin rest. A keyboard lay in front of the person on the table.

3.4 SUBJECTS

Ten persons, 8 men and 2 women, participated in this experiment, aged 25 – 55 years, median equal to 28.5 years. The participants had varying background, including technicians, research students and lecturers. All the subjects were paid for their participation. The subjects had normal vision, either uncorrected or when corrected

3.5 METHOD

3.5.1 Experimental design

The experiment was conducted in two sessions. At each session one presentation level with different image codings was shown. At each session the participant was shown the reference images. Ten training trials were given, but only at the first session. Next a determination of thresholds according to the Method of constant stimuli (Gescheider, 1985)¹⁰ was conducted. Then category ratings were obtained for the reference images, in groups and separately for each image. The second session was conducted after a break of about 10 minutes, except for one subject who had the second session on a different day. The presentation of the images was randomized for each person, both for the reference images and for the test images. Half of the subjects had one presentation level presented at the first session, while the other half had the other presentation level presented at this session.

3.5.2 Procedure

The participant was introduced to the experiment and personal details were recorded. A visual test with a so-called Dial-a-Chart from the R. H. Burton Company at the same distance as the monitor used, 56 cm, was then performed. The person

was then shown the reference images. He or she was told that these should be compared to the images presented during the experiment. If the image was perceived as worse than the reference image, the person should press ‘Y’ on the keyboard in front of him. If not, he was asked to press ‘N’. Each image sequence was shown for 3s and the inter-stimulus interval was at least 1s, but the next image was not presented before the person had given his response.

When the images had been presented the person performed the rating on the category scales. He or she should first do this for the entire reference image sequences in a group, and then for each reference image presented individually. When this was completed, a break was made, upon which the second presentation was shown to the subject. Each session took about one hour, thus in total about two hours for every participant.

The category scales used were named in Swedish, but the English translations are “Blockiness”, “Noise”, “Blur” and “Total impression”. Each one was graded from 0 to 10, with numerals shown and with verbal descriptions at the numerals 1, 3, 5, 7 and 9. A low number indicates a good measure of image quality and vice versa for a high number. Two of the category scales with the English translations are shown in Figure 5. The person could mark his opinion anywhere on the line for a category scale.

Blockiness was described as how large or how many rectangles, that the person thought the images could be divided into. *Noise* was described, somewhat tautologically, as how much noise that the person considered existed in the image. *Blur* was described as how clear and distinct that the person considered the images to be. *Total impression* was the total impression of the quality of the images.

The judgements on the category scales were only done at the 5% packet loss level. The participant was first requested to give a joint verdict of all the 6 scenes jointly, and then for each of the scenes separately.

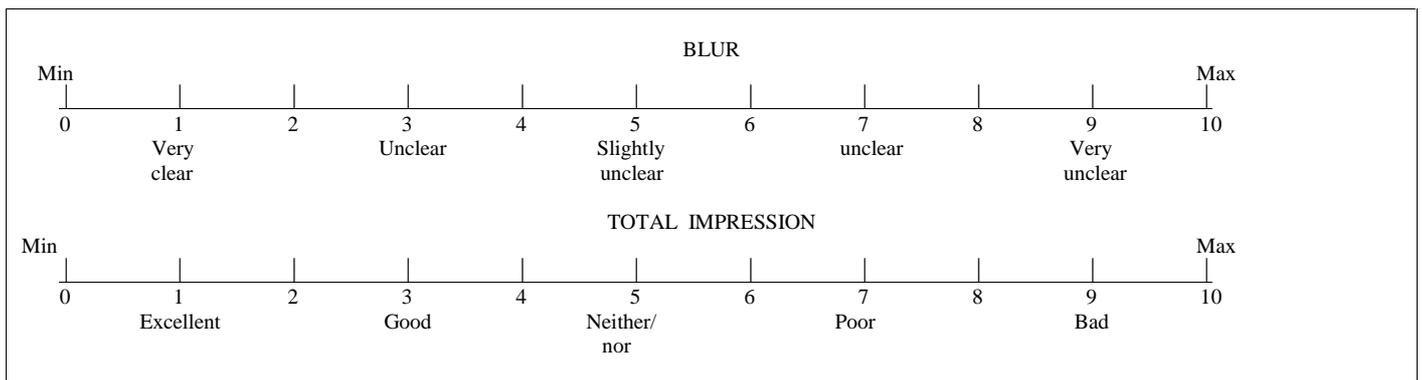


Figure 5: Two of the category scales used in Experiment 2.

4 RESULTS

4.1 ANALYSIS OF VARIANCE OF OBSERVERS DATA

In order to avoid the dependence of variance on the mean for results involving proportions, we transformed the proportion of yes-answers by $f(p) = 2\arcsin \sqrt{p}$, where p is the proportion yes-answers (see e.g. Howell, 1997, p.328)¹¹. In general, this transformation stretches out both tails of the distribution relative to the mean. An analysis of variance was performed for the transformed detection values of the participants. A mixed model was assumed, where the subjects are considered as the random factor. As a criterion of a significant effect we chose $\alpha=0.05$. The main effects for Scenes, S, and for Packet Loss, L, were significant, $F(5, 45)=6.96$ and $F(6, 54)=115.04$, respectively, while that for Presentation level, P, $F(1, 9)=0.35$ was not. The interactions of Scenes with Packet Loss, S*L, and that of Scenes with Presentation Level, S*P, were also significant, $F(30, 270)=4.12$ and $F(5, 45)=3.33$, respectively. The interaction of Packet loss with Presentation level, L*P, as well as the third order interaction of Scenes with Packet Loss and Presentation level, S*L*P were also significant, $F(6, 54)=3.13$ and $F(30, 270)=2.60$, respectively. However, Packet Loss and Scene main effects are still significant, but none of

the interactions are significant when we use the more conservative test based on the $F(1,9,95)=5.12$, as suggested by Winer (1962)¹². A similar analysis, for the non-transformed values was also done, with the same effects being significant.

The results shows that there were significant effects for the Loss variable, which of course was expected. More interesting is that there is an interaction effect of loss with presentation, L*P, i.e. the effects of the packet losses were different for the two presentations. The effects for Scenes at different levels of Presentation level, i.e. S*P, may explain the significant differences between the scenes, although the main effect of the two presentation levels, P, was not. As mentioned, another interaction effect with presentation level, namely that with packet loss, L*P, was also significant. These effects are illustrated in Figure 6.

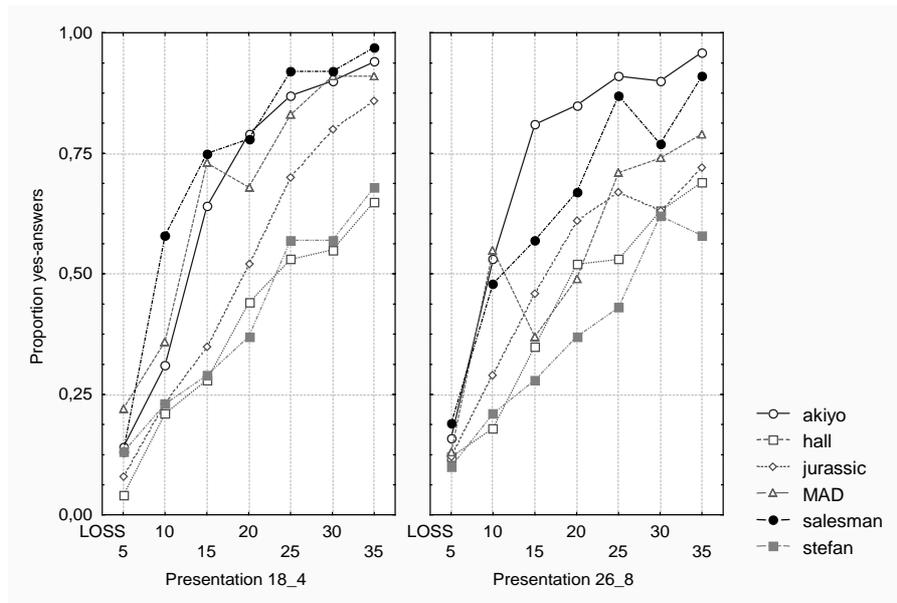


Figure 6: The mean effects of packet losses for the two presentation levels for the six scenes

4.2 MODEL PREDICTIONS OF DETECTION

One of the aims of this study was to compare the empirical threshold values with those estimated by the model. The 50% detection threshold values for each subject and each image and presentation level was determined by fitting a polynomial of the second order to the data, and finding the packet loss corresponding to the 50% detection value. The average threshold for each scene was only calculated for those subjects who had a 50% point on the resulting second-degree polynomial. The average thresholds for all thresholds of the subjects together with the predicted values given by the model are shown in Figure 7. The thresholds of the model were estimated by computing the model responses between the sequences containing packet loss distortions and the sequences containing H.263 coding distortions but no packet losses. A line was fitted between the model values in a logarithmic scale and the packet loss frequencies of the actual loss.. The thresholds were then estimated as the packet loss difference from the 5% level, which gave a model difference of 1 jnd. The correlation between these values of the model and the average empirical thresholds was 0.66 for presentation level 18_4, 0.80 for presentation level 26_8, and for both presentation levels 0.56 together. This illustrates a fair ability of the model to predict the differential effects of distortion for the different images.

** This computation might be problematic if the luminance becomes zero, but the lowest luminance used in this experiment was 0.1 cd/m^2 .

Besides the spatio-temporal model used in the present experiment to determine thresholds, other methods are possible, e.g. the contrast energy as mentioned above, but also the actual number packet losses and the PSNR-value.

The threshold values based on the PSNR values were calculated. The resulting values are also shown in Figure 7. The correlation, between the PSNR and the average thresholds, was 0.45 for presentation level 18_4, 0.62 for presentation level 26_8 and 0.47 for both presentation level together illustrating a slightly better ability of the model for predicting the empirical values. The main difference in correlation between the model and PSNR comes mainly from larger spread in the PSNR thresholds. This was confirmed by calculating the differences between the correlation coefficients by using the r' -transform of Fisher and testing the resulting Z-statistic¹². The z-value for the difference between the correlations for the model and PSNR for all the motives was $z = 0.25$, for the 18_4 scenes $z = 0.37$ and for the 26_8 scenes $z = 0.45$. None of these values are significant at $\alpha=0.05$.

Computing the model response between a distorted sequence and the undistorted original will give an estimate of the strength of the distortion. This was done for all the different scenes, distortion levels and versions. For details of the involved calculations see Ahumada et al (1998)³. In the current calculation the last square-root has not been used, keeping the model output in the form of difference energy.

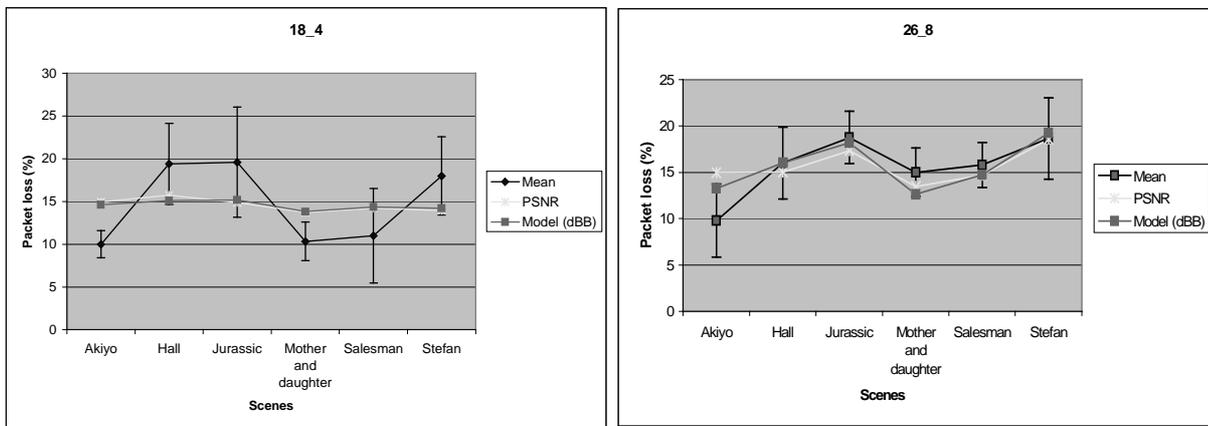


Figure 7: Empirical average 50% detection distortion thresholds for the subjects and model based thresholds.

4.3 ANALYSIS OF VERSION DEPENDENCE

The variable Packet loss varied with a probability from 5 to 35%. In order to vary accidental and strange instances of packet loss for a certain scene, we had 5 different versions of each scene at each packet loss level and each presentation level. However, since the variable is based on expected probability, a certain version may contain, for example, a higher amount of distortion than another version with a higher probability of packet loss. One way to get a measure of the actual extent of

the distortion is to measure the contrast energy, by computing $E_c = A \cdot t \cdot \sum_{x,y,t} c(x,y,t)^2 \text{ [rad}^2 \cdot \text{s]}$, where A is the area of one pixel in degrees squared, t the duration of one frame in seconds. The sums are taken over all pixels and all the frames in the sequence. The contrast of the distortion is estimated by $c(x,y,t) = L_d(x,y,t)/L_o(x,y,t) - 1^{\dagger\dagger}$, where L_d is the luminance for the distorted sequence and L_o is the luminance for the original undistorted sequence.

This was done for each presented version of every image. The average proportions of yes-answers for the ten subjects were then computed for each version. The correlation between the dependent variable Average proportion Yes-answers and Packet loss, Contrast energy in decibel units and Contrast energy in linear scale were 0.74, 0.26 and 0.24, respectively.

^{††} This computation might be problematic if the luminance becomes zero, but the lowest luminance used in this experiment was 0.1 cd/m^2 .

A multiple regression with the number of yes-answers of all the subjects as the dependent variable and scene, presentation level, scene version, expected packet loss, contrast energy in Decibel and in linear scale, actual packet loss, model energy and Peak Signal to Noise Ratio (PSNR) see Equation (3) **Error! Reference source not found.** The result is presented in . B and Beta are the non-standardized and the standardized regression coefficients, respectively.

	Beta	St. Err. of Beta	B	St. Err. of B	t(410)	p-level
Intercpt			-55.75	10.70	-5.21	<0.001
Scene	0.29	0.08	0.10	0.03	3.61	<0.001
Presentation level	0.16	0.06	0.19	0.07	2.56	0.011
Version	-0.02	0.03	-0.006	0.01	-0.61	0.54
Expected packet loss	0.12	0.07	0.007	0.004	1.59	0.11
Contrast energy (dB)	0.37	0.08	0.03	0.01	4.76	<0.001
Contrast energy (linear)	-0.07	0.04	-0.0006	0.003	-1.87	0.06
Actual packet loss	0.78	0.08	4.87	0.52	9.38	<0.001
Model energy	0.47	0.06	0.14	0.19	7.64	<0.001
PSNR	0.93	0.07	0.21	0.02	12.52	<0.001

Table 2: Summary of multiple regression on the average yes-answers

The R-correlation was 0.86 and its square, R^2 , showing the explained variance, was 0.74. As can be seen, the Actual loss parameter is the most significant parameter. Contrast energy apparently does not play the most important role for the detection of disturbances in the presented images when seen as a total. The actual packet loss and the PSNR measure are in this analysis the most important parameters.

$$PSNR = 10 \log_{10} \left(\frac{255^2}{MSE} \right) \quad MSE = \frac{1}{N \cdot M \cdot K} \sum_{n=1}^N \sum_{m=1}^M \sum_{k=1}^K (D_{nmk} - O_{nmk})^2 \quad \begin{array}{l} D = \text{distorted sequence} \\ O = \text{original sequence} \end{array} \quad (4)$$

To see how these measures compare to the model based values the correlation between the empirical values and the theoretical were computed for each scene and presentation level. In addition, the correlations with the expected loss frequencies were also computed. Each value is based on 35 observations. The PSNR has negative correlation as its value decreases with increasing distortions. The resulting correlations between the dependent variable and the various physical measures are shown in Table 3.

Scene	Pres. level	Expected loss	Actual loss	Contrast energy, dB	Contrast energy, lin	Model	PSNR (abs. values)
Akiyo	18_4	0.90	0.89	0.88	0.76	0.93	0.94
Akiyo	26_8	0.79	0.80	0.90	0.68	0.83	0.84
Hall	18_4	0.84	0.87	0.76	0.76	0.88	0.90
Hall	26_8	0.83	0.91	0.76	0.68	0.92	0.93
Jurassic	18_4	0.93	0.93	0.89	0.86	0.93	0.92
Jurassic	26_8	0.84	0.85	0.73	0.48	0.90	0.90
Mother and daughter	18_4	0.87	0.87	0.92	0.89	0.92	0.91
Mother and daughter	26_8	0.60	0.82	0.86	0.80	0.88	0.86
Salesman	18_4	0.83	0.83	0.89	0.85	0.89	0.89
Salesman	26_8	0.81	0.83	0.90	0.88	0.90	0.90
Stefan	18_4	0.92	0.92	0.88	0.91	0.90	0.88

Stefan	26_8	0.87	0.90	0.89	0.89	0.91	0.88
--------	------	------	------	------	------	------	------

Table 3: Correlations between the mean of yes-answers of the subjects to physical values of distortion of the images.

The table shows that when the image sequence is held constant the model based values correlate slightly better with the subject's judgments than do the loss measures or the contrast energy measures. However, the PNSR values appear to be on par with the model values. These correlations are based on the whole range of the responses of the subjects, whereas the threshold values only look at one point of the scale for the determination. Since two curves may have a high covariation, while also being far apart, the predictions of the model needs only to be close to the empirical responses at the vicinity of the threshold. Furthermore, the correlations of the values for the scenes with the predictions of the model tell us more of the success of the model than that for a single scene.

4.4 CATEGORY SCALES

The values for each of the four category scales were analyzed by analysis of variance, mixed model with subjects as the random factor. The main factors in each analysis were Subjects, Presentation level and Scenes with degrees of freedom of 9, 1 and 5 respectively. The interaction Presentation Level*Scenes thus had 5 degrees of freedom. The F-ratios of the analysis for the category scales is shown in Table 4.

	Blockiness	Noise	Blur	Total Impression
Presentation level, df=1	2.65	0.51	7.56, p>0.05	3.95
Scene, df=5	5.71, p<0.05	5.19, p<0.05	4.40, p<0.05	5.67, p<0.05
Presentation level*Scene, df=5	0.33	0.43	1.10	0.70

Table 4: The F-ratios resulting from the analysis of variance of the category scales for three of the effects.

The average values for all subjects for the scenes at the two presentation levels are shown in Figure 8. We here include the judgments for the scenes judged as a group, although this judgment was not included in the analysis of variances. It is e.g. apparent that the Mother And Daughter scene was most sensitive to disturbances at both presentations. This is most evident from the scale named Total impression. We believe that this scale should be seen as the most important criterion for how the persons perceive the images. The tennis player called "Stefan" has a low Total impression on both presentation levels. In general, the values of the 26_8 presentation are higher, i.e. a lower image quality, than the 18_4 presentation.

One sees a difference in the values afforded the images, where the 26_8 generally got worse impressions than the 18_4 presentation level. However, the difference between the two presentation levels was only significantly different from each other on the Blur scale. One may observe that the Total impression is judged worse than the other scales. A further investigation could find the psychological scales that constitute the Total impression.

The correlations between the thresholds and the judgments for the category scales for all the subjects (n=102 valid cases) showed that the correlation between the empirical threshold and Blockiness, Noise, Blur and Total impression were -0.145, -0.106, -0.067 and -0.002, respectively. We interpret the low correlations as indicating that the two dependent variables refer to two different psychological processes, one concerned with detection of distortions or disturbances, the other with overall quality judgments.

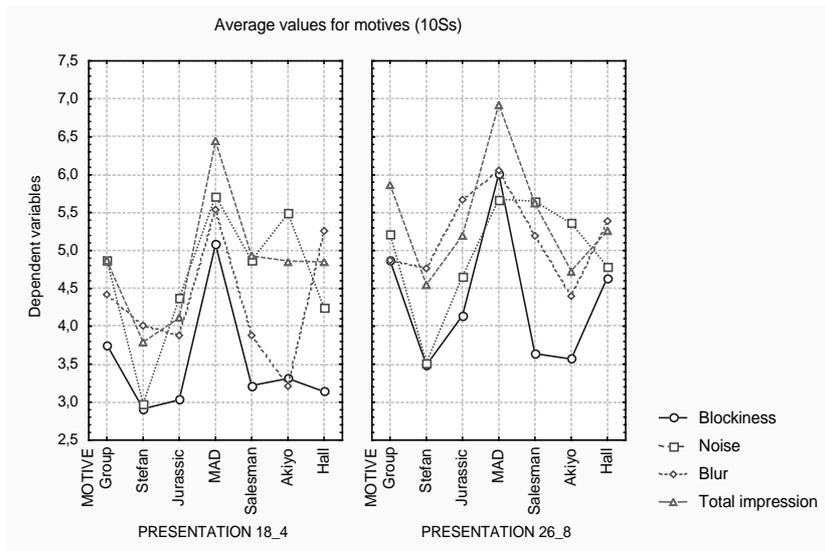


Figure 8: The mean judgments of the subjects in Experiment 2 for the scenes, as a group and separately, for the four category scales, at both presentation levels.

5 DISCUSSION

For Experiment 1 the spatio-temporal model was a good predictor of distortion detection thresholds for video image sequences. For Experiment 2 the model failed to predict for the better quality presentation level, i.e. 18_4, which of the image sequences would be the most affected by packet loss, but it performed better for the worse quality level, i.e. 26_8. The method of constant stimuli with a remembered reference gave consistent results. Category scaling gave interesting data.

We had wanted to investigate if it was possible to use a method for quality surveillance, where customers use a reference for determining the quality. However, it is also possible that the participants used an inner reference that was not dependent on the earlier presentations. The remarks of the subjects after the study indicated that they had been unable to use the reference images presented at the beginning of each session. In order to study if the subjects used the inner reference or not, one would need to conduct a study for this particular issue.

The results indicate that the sensitivity to the distortions is higher for the so called head-and-shoulder image sequences compared to the others. These scenes contain very little movement that could mask the errors. For instance, 'Stefan' have been judged as having the best overall quality as reference image, see Figure 8, and shows very low sensitivity for increasing rates of packet losses, see Figure 6. This is also the scene, which contain the most movements. Another important aspect is the sensitivity of the human visual system for disturbances located in faces. Attempts have been made to account for this when coding, that is to first locate the facial region of a human and then allocate more bits to this region (Daly et al (1999))¹³. From Figure 7 it is apparent that an additional packet loss of 5 % was sufficient for some scenes and 15% is sufficient for all scenes for 50% detection of a distortion.

We computed some other image based measures on the image sequences, contrast energy, spatio-temporal model energy and peak-signal-to-noise ratio. The PSNR values predicted the differences among the image sequences almost or about as well as the spatio-temporal model.

One difference in the two experiments is that the three subjects in the first study were all active vision researchers, whereas the subjects in the second study had more varied backgrounds. Another important difference is that in Experiment 1 the thresholds have been computed directly from the difference between the undistorted original and the distorted sequences. In Experiment 2, the thresholds are based on the difference between the distorted reference sequences and the more distorted sequences. For reasons of adequate calculations due to the character of packet losses, the model looked at the difference between the coded original scene and the distorted scenes. The findings support the conclusions that for multidimensional distortions, the model predictions do not always fit the mean judgments of observers. These conclusions corroborate those

of Ahumada and Null (1993)¹⁴ and the category scaling in the present study also illustrate that, once distortions are supra threshold, different subjects can weigh different dimensions differently so that no single measure of image quality is possible.

The results in Experiment 2 for the detection values and for the category scales give different results for the scenes. The Mother and Daughter scene is e.g. not much different from the average of the other scenes regarding the detection values, but much different on the category scales. We believe that this mirrors the existence of two psychological processes, one is used to detecting differences or distortions, while the other is used to form a general impression of the image. One may e.g. detect a distortion in an image, but not feel that it matters so much for the quality impression and vice versa. This is maybe similar to the distinction between local and global psychophysics (Martens and Boschmann, in press)⁶, where the former refer to methods aimed at determining detection and discrimination thresholds, while the latter refer to what are called supra-threshold measurements. Roufs and Boschmann (1991)¹⁵ advocate the relevance of distinguishing between performance- and appreciation-oriented perceptual quality measures. Visual comfort is measured more adequately by numerical category scalings, than by basic psychophysical or physiological methods. We therefore believe that our results illustrate the need for both types of measures for studies of image quality.

The predictions of the model were for some of the conditions somewhat close to the empirical values. The conditions that may make it difficult for predictions are the quality level of the original scene, the qualitative content of the scene and the task of the observer. The benefit of the model in relation to simpler measures such as the PSNR ought to be further investigated.

6 ACKNOWLEDGMENTS

We are grateful to advice provided us by prof. Sture Eriksson, Uppsala University and by prof. Birgitta Berglund, Stockholm University. Mr Börje Andrén was helpful, especially in conducting photometric measurements. The observers' contributions are gratefully acknowledged. This work was supported by the companies Telia Research AB, Ericsson SAAB Avionics AB, CelsiusTech Electronics AB, FMV, and NUTEK (Swedish Board for Industrial and Technical Development).

7 REFERENCES

1. A. B. Watson, J. Hu, J. F. McGowan III and J. B. Mulligan, "Design and performance of a digital video quality metric", Human Vision and Electronic Imaging IV, Proceedings 3644, SPIE: Bellingham, WA, pp. 168–174, 1999.
2. S. Winkler, "Perceptual distortion metric for digital color video", Human Vision and Electronic Imaging IV, Proceedings 3644, SPIE: Bellingham, WA, pp. 175–184, 1999.
3. A.J. Ahumada, Jr, B.L. Beard and R. Eriksson. Spatio-temporal image discrimination model predicts temporal masking functions. In B. Rogowitz & T.N. Pappas (eds.) Human Vision and Electronic Imaging III, Proceedings 3299, SPIE: Bellingham, WA, pp. 120–127, 1998.
4. K. Brunnström, R. Eriksson, B. Schenkman and B. Andrén, "Comparison of predictions of a spatio-temporal model with responses of observers for moving images". Technical Report, TR-338, Inst Optical Research, Kista, Sweden, 1999.
5. G. Borg, "Psychophysical judgment and the process of perception". In H.-G. Geissler and P. Petzold (eds.) Psychophysical judgment and the process of perception. VEB Deutscher Verlag der Wissenschaft, 1982.
6. J.-B. Martens and M. Boschmann, The psychophysical measurement of image quality. In press.
7. Methodology for the subjective assessment of the quality of television pictures. Rec.ITU-R BT.500-7, International Telecommunication Union, 1974-1995.
8. J A J Roufs and M C Boschman, "Text quality metrics for visual display units: I. Methodological aspects". Displays, 18(1), 37-43, 1997
9. Video coding for low bit rate communication, ITU-T Recommendation H.263, International Telecommunication Union, 1998
10. G.A. Gescheider. Psychophysics. Method, Theory and Application. Hillsdale, NJ, Lawrence Erlbaum, 1985.
11. D.C. Howell. Statistical methods for psychology. Duxbury Press: Belmont, CA, 1997
12. B. J. Winer, *Statistical Principles in Experimental Design*. McGraw-Hill: New York, NY, 1962.
13. S. Daly, K. Matthews and J. Ribas-Corbera, "Visual eccentricity models in face-based video-compression", In B. Rogowitz & T.N. Pappas (eds.) Human Vision and Electronic Imaging IV, Proceedings 3644, SPIE: Bellingham, WA, pp. 152–166, 1999.

14. A. J. Ahumada, Jr. and C. Null, "Image Quality: A Multidimensional Problem" in A. Watson, ed., *Digital Images and Human Vision*. MIT Press: Cambridge MA, 1993.
15. J.A.J. Roufs and M.C. Boschman. "Visual comfort and performance". In J.A.J Roufs (ed.) *The Man-Machine Interface*. Macmillan, London, 1991.