EI 3959-18

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

Kjell Brunnström[a], Bo N. Schenkman[a], Albert J. Ahumada Jr.[b]

[a]ACREO AB, Electrum 236, SE-164 40 Kista, Sweden

[b]NASA 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 processing channels are used and masking rules are based on contrast gain controlTwo 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 evaluation with versions of H.263 compressed monochrome 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 image video quality was worse than the initially seen reference imagevideo. Two different quality levels of the compressed sequences were simulated. Category scales were used for further assessments. The scene-wise correlations between model and subjective data were high, but the model performance was comparable to that of other measures. Predictions of more general nature were not high or were predicted better by other measures.

1. 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 videoconferences, starts to become available for the ordinary customer. However, the quality is still poor, especially image quality for video conferences. This is due to the limited bandwidth and to the packet based transmission. The limited bandwidth will force the use of high compression rates and the packet based transmission gives very little control of the exact arrival time of the transmitted packets.This means that in a real-time application the delayed packets can be included or discarded upon arrival, but in either case, they will 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 involved for using this type of transmission. It could therefore be envisioned that a customer will be provided with a quality level for which they are prepared to pay for. We have in this study been focussed on the detection ofpoorer quality, given an already known level and in what way this could be predicted with a visual model.

One way of ensuring that a good or at least satisfactory quality of images over the Internet is obtained is by using a visual model with a reference image to survey the transmitted images. The reference image would correspond to a certain image quality. There have been many reports, e.g. at SPIE-conferences, in presenting efforts to include findings about the early-vision system into a computational model, that may be used 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. This is a spatio-temporal visual model that was presented earlier by Ahumada et al (1998)[3], 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. The image sequence discrimination model that we use models optical blurring and retinal light adaptation. The processing proceeds in two parallel channels, here called Magno and Parvo, responding principally to the high temporal and low spatial frequencies (i.e. transient channel) and the low temporal and high spatial frequencies (i.e. sustained channel), respectively. This simulates the separation of processes in the ganglion cells and in the Magno and Parvo structures in the Lateral Geniculate Nucleus. The masking rules are based on contrast gain control and are different for the two channels.

In this article we describe two experiments that was designed to study the performance of the model for a task that is representative of a video communication system. In the first experiment it was assumed that there was no transmission errors, but possibly limited bandwidth. The primary aim of Experiment 1 was to establish the validity of the model for detecting disturbances in video images.

In the second experiment it was assumed that the transmission errors in the form of lost packets could occur, e.g. as on the Internet. In the first Experiment below we used a 2-alternative forced choice method with a reference shown together temporally with the presented image. This is a common psychophysical method when studying detection. However, in a real situation, a user may not have access to a reference image. A user of Internet services may not know of the image quality of the original or he or she may have

seen it some time ago, and then makes a comparison to an image stored in his or her memory. We therefore changed the method in Experiment 2 to incorporate this memory aspect. The test person here had to compare the image presented to the remembered image in his or her memory. The first aim of Experiment 2 was to see if this memory method was a feasible method for understanding the image quality problems of transmission systems

It would be useful to find a model that could predict the eventual judgments of users concerning image quality. One such model is the spatio-temporal model mentioned above. However, 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 both by S.S. Stevens and G. Ekman (see Borg, 1982)[5]. They only approved of methods with ratio scalings. In recent psychophysical literature (cf. Martens and Boschmann, in press)[6] the use of category scales is advocated 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

1. Experimental method

Five monochrome image sequences were used to generate pairs to be discriminated. 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. Results

The thresholds for each subject at each of the image sequences were calculated. Analysis of variance was computed also including the predictions of the model, giving 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 1Figure 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 for the mean of the three real subjects.

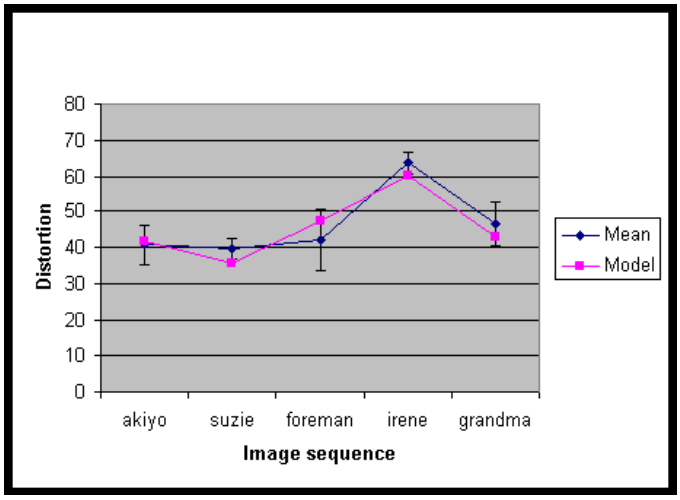

Figure 1: The mean thresholds of the subjects in Experiment 1 and the predicted thresholds of the model. The vertical bars for the means show 95% confidence intervals.

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 Brunnström, 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 must be viewed with most caution. In order to confirm the results, Experiment 2 with more subjects was conducted, where also a greater variety of image sequences with different contents and characteristics was used.

In real time video transmission today on the Internet, compressed images are transmitted as packets. We were therefore interested in seeing how the model would cope with such an application. 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

1. Method

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. Then the determination of thresholds according to the Method of constant stimuli (Gescheider, 1985)[10] was conducted. When this was completed category scales was shown 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.

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 for each session, thus in total ca 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 3Figure 2.

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 person could mark his opinion anywhere on the line for a category scale.

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.

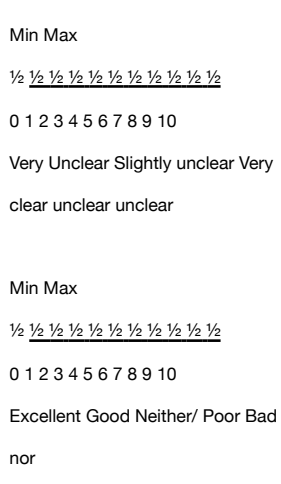

Figure 32: Two of the category scales used in Experiment 2.

2. SCENES/stimuli

Images, i.e. scenes were compressed according to the H.263 standard using two layers, one base layer and one improvement layer. The improvement 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 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 improvement 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 was 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 5Figure 3 and 85 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 4Figure 4**Error! Reference source not found.**.

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 53: 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 4: 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.

Figure 85: 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. Room conditions

The participant sat in a small chamber with gray homogenous 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 as well as the table in front of the person was also covered with the gray cloth.

4. 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.

5. 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..

4. RESULTS

1. Detection values

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) = 2arcsin Ö p, where p is the proportion yes-answers (see e.g. Howell, 1997, p.328). 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 a =0.05. In the model chosen and since there are no replications we could not test the effects of the random variable, i.e. subjects, nor its interactions with the other variables. 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. A similar analysis, for the non-transformed values was also done, with the same effects being significant.

The similarities for the two presentation levels are shown in Figure 11 for the non-transformed detection values. It is likely that the similarities for the two levels point to similar underlying psychophysical processes for the two scenes presented at the two levels, since differing appearances would point to dissimilar processes.

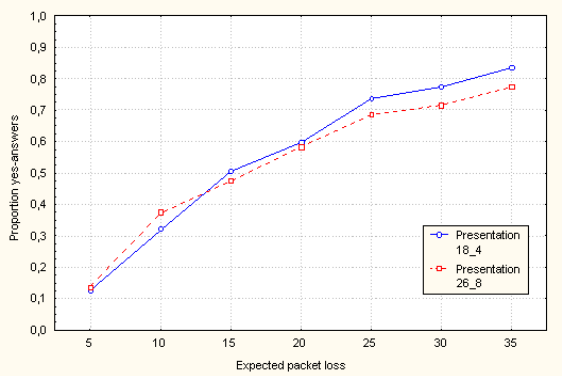

Figure 76 The mean proportion of detection of distortion for the two presentation levels

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 8Figure 7.

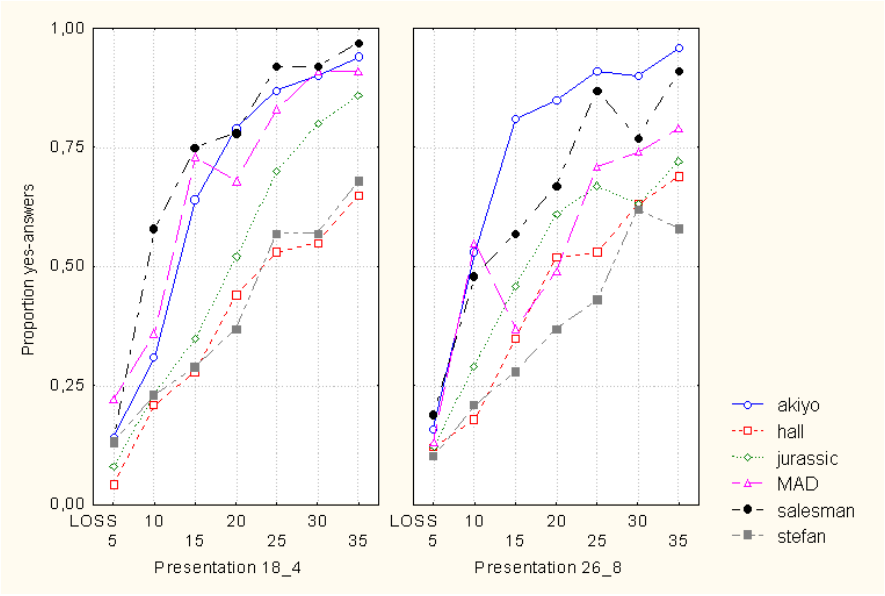

Figure 87The mean effects of packet losses for the two presentation levels for the six scenes

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 $E_c = A \sum_{x,y,t} c(x,y,t)^2 \left[\mathrm{rad}^2\right]$, where A is the area of one pixel in degrees, 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 $c(x,y,t) = L_d(x,y,t) / L_o(x,y,t)$, 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 on decibel units and Contrast energy in linear scale were 0.74, 0.26 and 0.24, respectively.

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). The result is presented in Table 1Table 1. 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 1.:. Summary of multiple regression on the average yes-answers

The R-correlation was 0.86 and its square, R2, showing the explained variance, was 0.74. As can be seen, the Actual loss parameter is still 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.

2. Model COMPARISONS

One of the aims of this study was to compare the empirical threshold values with those estimated by the model. Since the model uses the concept of an average observer, we used the mean values of all the subjects. The mean threshold values of the subjects for a detection of 50 percent for each scene was calculated on basis of the mean values for all the subjects, as shown in Figure 8Figure 7 above. The estimated value was determined by fitting a polynom of the second order to the data, and finding the packet loss corresponding to the 50% detection value. Each threshold is based on 70 observations. The empirical thresholds values determined for the different scenes was compared to the model-based threshold values. . The correlation coefficient for the resulting data set of n=12 was r=0.17.

However, this way of computing thresholds over all the subjects does not take into account that some subjects for some scenes never saw a worse distortion, while some subjects always saw a distortion for some scenes. We therefore computed individual thresholds for each subject and scene and each presentation level. The average threshold for each scene was only calculated for those subjects, who had a 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 8. The correlation was –0.27 between the model and the average thresholds.

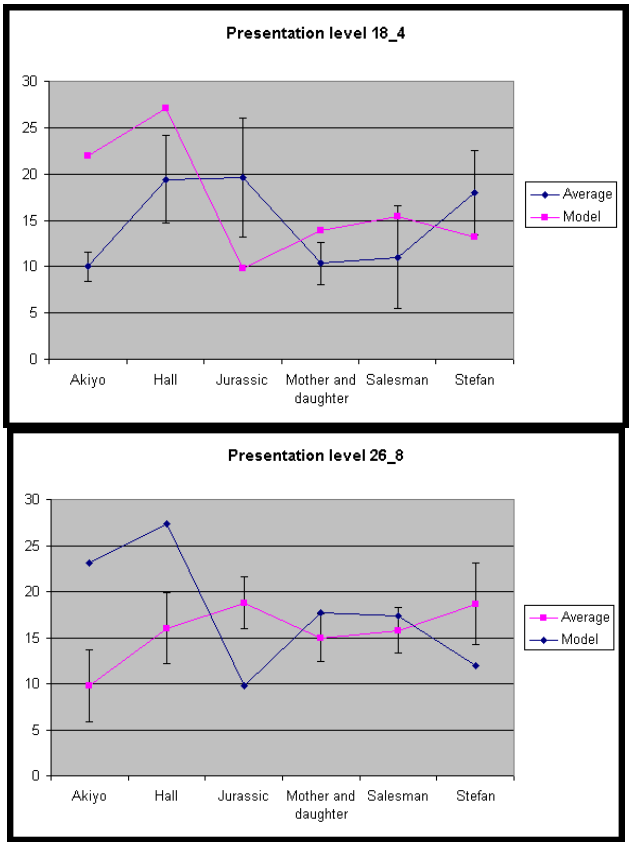

Figure 108: Empirical average thresholds based on subject's thresholds and model based thresholds, both showing a packet loss resulting in a detection of 50% probability.

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.

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 Signal-to-Noise ratio (PSNR), see Equation . 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 3Table 2.

$$PSNR = 10\log_{10}\left(\frac{255^2}{MSE}\right) \qquad MSE = \frac{1}{N \cdot M \cdot K}\sum_{n}^{N}\sum_{m}^{M}\sum_{k}^{K}(D_{nmk} - O_{nmk})^2$$

D = distorted sequence

O = original sequence

* MERGEFORMAT ()1

Scene	Pres. level	Expected loss	Actual loss	Contrast energy, dBB	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 32: Correlations between the mean of yes-answers of the subjects to physical values of distortion of the images.

The table shows e.g. that 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.

3. 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 5Table 3.

	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 53: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 11Figure 9. 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.

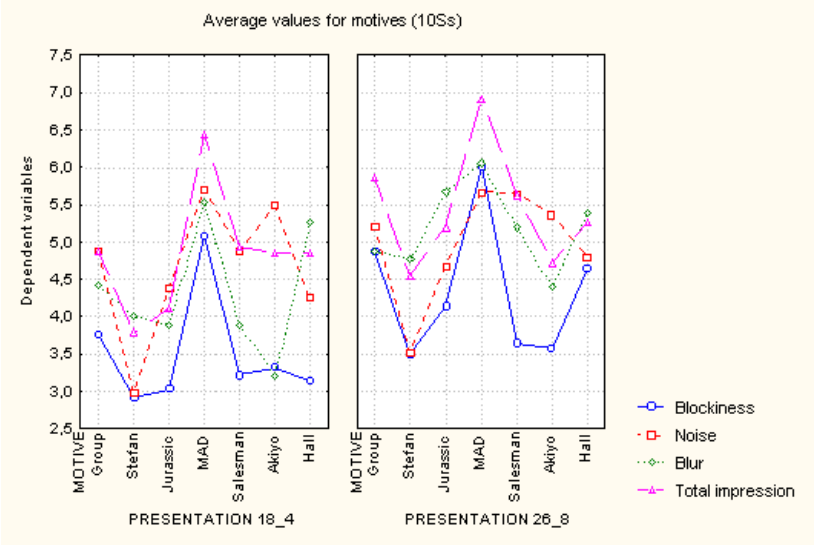

Figure 119: The mean judgments of the subjects in Experiment 2 for the scenes, as a group and separately, for the four category scales, at both presentation levels.

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.

5. DISCUSSION

Experiment 1 had shown that the spatio-temporal model was a good predictor of the detection of distortion for video images. Experiment 2 showed that this or higher correlation could also be reached by using other and simpler measures such as the amount of packet lost or the PSNR. The use of a remembered reference was possible, but it is doubtful what information or cues the participants actually used. The use of category scales was a useful method.

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. The apparent similarity of the empirical curves for the two presentation levels, see Figure 6 could indicate that this could be a feasible method. However, it is also reasonable that the similarity derives from the use the participant made of an inner reference that was not dependent on what had been presented earlier. 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. The similar appearances of the curves do however, we think, point to underlying similar psychological processes. 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 are 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 11Figure 9, and shows very low sensitivity for increasing rates of packet losses, see Figure 8Figure 7. 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))[11]. From Figure 8 it is apparent that an additional packet loss of 5 % was sufficient for some scenes and 15% is sufficient for all scenes for detection of a distortion.

An interesting observation is the strong connection between the average score of yes response for the different scenes and various types of descriptions of the influence of the packet loss. The expected packet loss, correlates well with the average yes score, apart from the Mother-and-Daughter scene. In this case there is a mismatch between the expected number of packet losses and the actual number. Not surprisingly, the actual loss correlates better, as shown in Table 3Table 2. We computed also some image based measurers on the image sequences, contrast energy, spatio-temporal model energy and signal-to-noise ratio. The model energy and the PSNR correlates equally well. Further refinements of the model could be accomplished by combining a number of these and other measures into a statistical model, or by changing the model, e.g. by introducing other parameters.

A threshold value could be defined on 50 % yes frequency, which could be used as a packet loss level were an increased distortions will most likely be noticed by an observer. These levels are very similar for the two presentations, but differs between the different scenes. However, these values were not well predicted by the model in the second experiment. The differences to the first experiment in this regard could be due to that the three subjects in the first study were all active vision researchers, whereas the subjects in the second study had a more varied background. 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 of the model responses computed between the reference sequences and the original, and the more distorted sequences and the original. To improve these predictions, the model need to extract a higher level description of the strength of distortion and compare that with the strength of the distortion of the other sequence.

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 show that one should use both types of measures for studies of image quality.

6. ACKNOWLEDGMENTS

We are grateful to advice provided us by prof. Sture Eriksson, Uppsala University and by prof. Birgitta Berglund, Stockholm University. Mr Börje Andren was helpful, especially in counducting 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.
1. 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.
1. 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.
1. J.-B. Martens and M. Boschmann, The psychophysical measurement of image quality. In press.
2. Methodology for the subjective assessment of the quality of television pictures. Rec.ITU-R BT.500-7, International Telecommunication Union, 1974-1995.
1. J A J Roufs and M C Boschman, "Text quality metrics for visual display units: I. Methodological aspects". Displays, 18(1), 37-43, 1997
2. Video coding for low bit rate communication, ITU-T Recommendation H.263, International Telecommunication Union, 1998
1. G.A. Gescheider. Psychophysics. Method, Theory and Application. Hillsdale, NJ, Lawrence Erlbaum, 1985.
2. 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.
3. D.C. Howell. Statistical methods for psychology. Duxbury Press: Belmont,CA, 1997.
1. Video coding for low bit rate communication. ITU-T Recommendation H.263, International Telecommunication Union, adopted 02/1998.
1. 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.