Comparison of the predictions of a spatiotemporal model with the detection of distortion in small moving images

Kjell Brunnström, MEMBER SPIE
Bo N. Schenkman
ACREO AB
Electrum 236
164 40 Stockholm
Sweden
E-mail: Kjell.Brunnstrom@acreo.se
Bo.Schenkman@acreo.se

Abstract. The image sequence discrimination model we use models optical blurring and retinal light adaptation. Further, two parallel channels, sustained and transient, with different masking rules based on contrast gain control, are also used. Performance of the model is 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 constitute pairs of noncompressed and compressed images to be discriminated with a two-alternative-forced-choice method together with a staircase procedure. The discrimination thresholds for each subject are calculated. Analysis of variance shows that the differences between the pictures are significant. The model threshold is close to the average threshold of the subjects for each picture, and the model thus predicts these results quite well. In the second study, the effect of transmission errors on the Internet, i.e., packet losses, is tested with the method of constant stimuli. Both reference and comparison images are distorted. The task of the subjects is to judge whether the presented video quality is worse than the initially seen reference video. Two different quality levels of the compressed sequences are simulated. The differences in the thresholds among the different video scenes are to some extent predicted by the model. Category scales indicate that detection of distortions and overall quality judgements are based on different psychological processes. © *2002 Society of Photo-Optical Instrumentation Engineers.*
[DOI: 10.1117/1.1431551]

Subject terms: video; image quality; spatiotemporal; vision model; H.263; packet loss; Internet.

Paper 200261 received June 30, 2000; revised manuscript received Mar. 14, 2001; accepted for publication Aug. 31, 2001.

1 Introduction

The Internet provides a huge infrastructure for connecting people in inexpensive ways over large distances. Services such as telephony and videoconferences are becoming available to the ordinary customer. However, the quality is still poor, especially image* quality for videoconferences. This is caused by 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 on 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 customers may then be provided with a quality level that they can afford. One approach to control the desired image quality level is to use a visual model to compare reference images of acceptable quality with the transmitted images. In this paper, we measure the detection of poorer quality by viewers and see whether a visual model can predict this detection.

There have been many reports of similar efforts to model the early vision system and use the model in technical applications. By an early vision system we mean the part of the visual system that is peripherally in the information process, and thus has not had much processing by the brain. Examples of models aimed at video applications are those presented by Watson et al.[1] and by Winkler.[2] This paper describes the success of such a model, a spatiotemporal model, in predicting the detection of image compression distortion in image sequences. We use the spatiotemporal visual model that was presented earlier by Ahumada et al.,[3] who evaluated its performance for contrast sensitivity and masking. Another study compared the predictions of the model with human performance of target detection in moving IR images.[4] One of our intentions in these experiments was to test this model for video applications. The model takes as input two luminance image sequences, applies the same processing on both of them, and gives as the output a single number, which is the predicted discrimination of the two image sequences in just-noticeable differ-

*We use the word ''image'' to denote an image sequence, moving image, or video, while nonmoving images are denoted by, e.g., ''still image'' or ''individual frame.''

ences. It has processing stages representing optical blurring and retinal light adaptation. The processing then proceeds in two parallel channels, one called Magno, responding to higher temporal and lower spatial frequencies (the transient channel), and one called Parvo, which is 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 that are based on contrast gain control are applied in the two channels. The model is a lumped parameter model, so that the low-pass filtering attributed to the optics of the eye is the combined effect of low-pass filtering at all stages.

In this paper, we describe two experiments designed to assess the utility of the model for predicting when observers detect video communication image 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 two-interval alternative-forced-choice method, with the test and reference sequences shown one after the other. 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 must 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 how different physical measures could explain the judgments of the observers.

One method of measuring the sensory experiences or psychological dimensions of people is to use category scales. These are scales that refer to a certain adjective or characteristic, for example, blur, ranging from a large value, to a small value, and the test person is asked to indicate where on the scale his or her experience or judgment falls. The use of category scales in psychophysics was criticized by Stevens and by Ekman (see Borg),[5] presumably because a category scale is not linearly related to a ratio scale. Stevens was mostly interested in general functions and wanted to compare different perceptual modalities with regard to functions and exponents. For this purpose, they only approved of methods resulting in ratio scales. Martens and Boschmann[6] advocate the use of category scales as a method of understanding image quality. The use of category scales with numerals is also included in assessment procedures for television pictures.[7] In the realm of visual display units, Roufs and Boschman[8] found that numerical category scaling offered a fast and efficient method for measuring the psychological attribute ''visual comfort.'' Roufs and Boschman said that this was the most relevant metric, since it reflected a combination of attributes that comprised the visual impression of the displayed text. The participants were also found to judge in a consistent way. 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 sequence pair presented, 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.[9] The task of the participant was to identify which of two sequences was distorted. The psychophysical method was two-alternative-forced-choice in combination with a staircase procedure adjusting the distortion level. There were three male participants with normal vision. The frame rate of the presentation was 30 Hz, giving a Nyquist frequency of 15 Hz. The observers sat between 50 to 60 cm from the monitor, a Sony Trinitron 20-in. screen. The screen resolution on this distance gave a spatial Nyquist frequency of about 9.5 cycles/deg. This experiment is described in further detail in Brunnström et al.[4]

2.2 *Results*

The just-detectable distortion thresholds were calculated for each participant for each of the image sequences. We also viewed the predictions of the model for a virtual subject, and these values were analyzed together with those for the real persons. A useful statistic is the F statistic, which is employed by the statistical method called analysis of variance. One can informally say that the F statistic tests the null hypothesis of no effects being present. If this is the case, then the expected value should be close to 1. The more that the value is different from 1, the less likely it is that the effect was caused by chance. The F statistic is obtained by dividing $MS_{treatment}$ by MS_{error} (MS stands for mean square), where the former gives an estimate of the population variance, but only if the null hypothesis is true, while the latter term always is an estimate of the population variance. If the null hypothesis is false, then $MS_{treatment}$ is different from MS_{error} and hence the F statistic is different from 1. The degrees of freedom for the F statistic is related to the investigated levels of the terms used in the numerator and the denominator, respectively. The fewer degrees of freedom, the higher the resultant F value must be to be significant at a certain significance level. Analysis of variance, using the participant 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 ''participants.'' These are both significant at $p=0.05$. The model response was close to the average of the participants (see Fig. 1). When computing *a priori* tests in the form of t tests for the difference

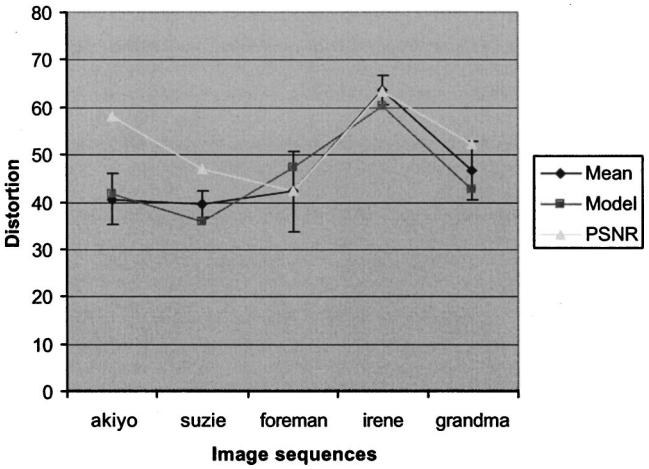

Fig. 1 Mean thresholds of the subjects in experiment 1 and the predicted thresholds of the model. The vertical bars for the means show 95% confidence intervals.

between the means, the threshold of the model did not differ significantly from that of the mean for the three participants for any of the five image sequences. The PSNR values were also calculated according to Eq. (1). As we can see, the fit between the mean empirical threshold values and the PSNR values is not as good as for the one between the threshold values and the model values.

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

where MSE is mean square error, D is the distorted sequence, and O is the original sequence.

2.3 *Discussion*

As we can see in Fig. 1, for three images, the predictions of the model were close to the mean of the three participants. For two of the images (''Suzie'' and ''Irene'') the fit was less good. The variance of these two images across the participants was lower than for the other three, which caused the predictions to fall on the border of the 95% confidence interval, although the deviation of the model mean from the empirical mean is not greater in actual values. Although more complex cognitive mechanisms may be important for some sequences, the model does predict the results quite well. Experiment 1 was planned to be a small pilot study and the results are 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 participants and a greater number and a greater variety of image sequences.

3 Experiment 2: Packet Losses

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.

3.1 *Method*

3.1.1 *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 all packets transmitted are retained. Six different scenes were compressed and the quantization parameters were varied for the two layers. For one set, these parameters were set to 18 and 4, while for the other they were set to 26 and 8. The first value refers to the quantization of the base layer and the second value to that for the enhancement layer. These two combinations can be called presentation levels, but a more correct name is probably compression strategies and they are subsequently denoted 18_4 and 26_8. The actual physical meaning of the parameter values is inherent to the coder used.

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

The reference images, i.e., scenes with packet loss probability of 5% are shown in Figs. 2 and 3 for the two compression strategy levels 18_4 and 26_8, respectively. (For copyright reasons the reference image for the scene ''Movie'' is not shown.) As a comparison, we have chosen to present for one image, ''Mother and Daughter,'' images at the 18_4 compression strategy for 20 and 35% packet loss (see Fig. 4).

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,'' ''Movie,'' and ''Stefan'' were chosen to represent images that could occur in a surveillance situation or in entertainment activities, e.g., on the Internet.

Each scene was shown for only 3 s to avoid any recency effects, which were discussed by e.g., Pearson[10] and de Ridder and Hamberg[11] for longer duration sequences. The quality of the last 10 to 20 s of a video will have largest effect of the judgment of an observer and the length of this effect is believed to be related to the length of human short term memory.

3.1.2 *Room conditions*

The participant sat in a small chamber with gray homogeneous cloth surfaces in front of, above, and to the sides of the person. The room illuminance on the screen was 96 lx,

Fig. 2 Tenth frame of the images "Akiyo," "Hall," "Mother and Daughter," "Salesman," and "Stefan" for the compression strategy 18_4 with a packet loss of 5%.

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.

3.1.3 *Apparatus*

The monitor used was an Eizo, 17 in., model T562-T with 800×600 pixel resolution. The maximum luminance level for gray level 255 was set equal to 100 cd/m^2. The resulting gamma function was measured with this original setting. This function was also used in the latter calculations of the model values. The screen refresh 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 number of pixels were 160 horizontally and 128 vertically, with frame rate of 15 Hz, giving a temporal Nyquist frequency of 7.5 Hz. The small size of the images was chosen, since most applications for packet-based transmission would probably use a small window on a screen. The participant sat at a distance of 56 cm from the screen, with his chin on a chin rest, giving a pixel size of about 2.4 arcmin, giving a spatial Nyquist frequency of

Fig. 3 Tenth frame of the images "Akiyo," "Hall," "Mother and Daughter," "Salesman," and "Stefan" for the compression strategy 26_8 with a packet loss of 5%.

Fig. 4 Tenth frame of the "Mother and Daughter" image for the compression strategy 18_4 for the 20 and 35% packet losses (left and right), respectively.

about 12 cycles/deg. A keyboard lay in front of the person on the table. Only monochrome images were shown to the observers.

3.1.4 *Observers*

Ten persons, 8 men and 2 women, participated in this experiment. Their ages were 25 to 55 yr with a median age equal to 28.5 yr. The participants had varying backgrounds, including technicians, research students, and lecturers. All the participants were paid for their participation. The participants had normal vision, either uncorrected or when corrected.

3.1.5 *Experimental design*

The experiment was conducted in two sessions. At each session one compression strategy with different image codings was shown. At each session the participant was shown the reference images. Ten training trials were given, but only in the first session. Next, a determination of thresholds according to the method of constant stimuli[12] 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 min, except for one participant 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 participants had one compression strategy presented at the first session, while the other half had the other compression strategy presented at this session. Each session took about 1 h.

3.1.6 *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 was performed at the same distance from the monitor (56 cm) as was used in the experiment. The person was then shown the reference images. He or she was told that the images presented during the experiment should be compared to it. 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 3 s and the interstimulus interval was at least 1 s, but the next image was not presented before the person had given his response.

When each image had been presented, the person performed the rating on the category scales. He or she should first do this for all the reference images as a group, and then for each reference image when presented individually. When this was completed, there was a break, after which the second presentation was shown to the person. Each session took about 1 h, thus in total about 2 h for every participant.

The category scales 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 for 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 Fig. 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 the person thought the images could be divided into. Note that this is a "strong" measure of blocking, requiring high levels of distortion. At lower levels of distortion, blocking can be perceived as sporadic edge and line breaks. *Noise* was described, somewhat tautologically, as how much noise that the person considered existed in the image. *Blur* was described as how unclear and nondistinct the person thought the images were, varying from very clear to very unclear. *Total impression* was the total impression of the image quality, varying from excellent to bad.

The judgments on the category scales were only done at the 5% packet loss level. The participant was first requested to give a verdict on all the six scenes together and then for each of the scenes separately.

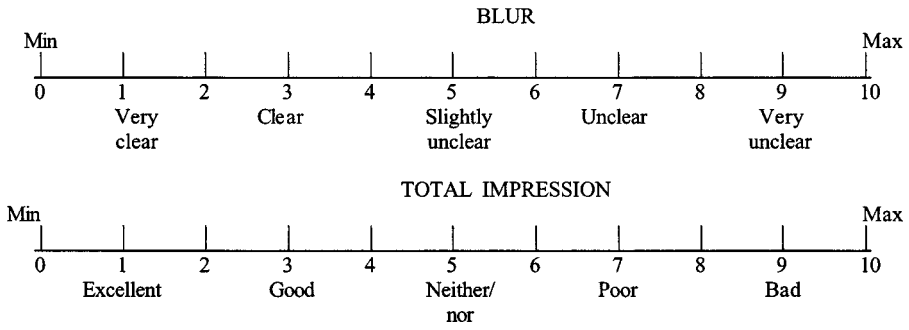

Fig. 5 Two of the category scales used in experiment 2.

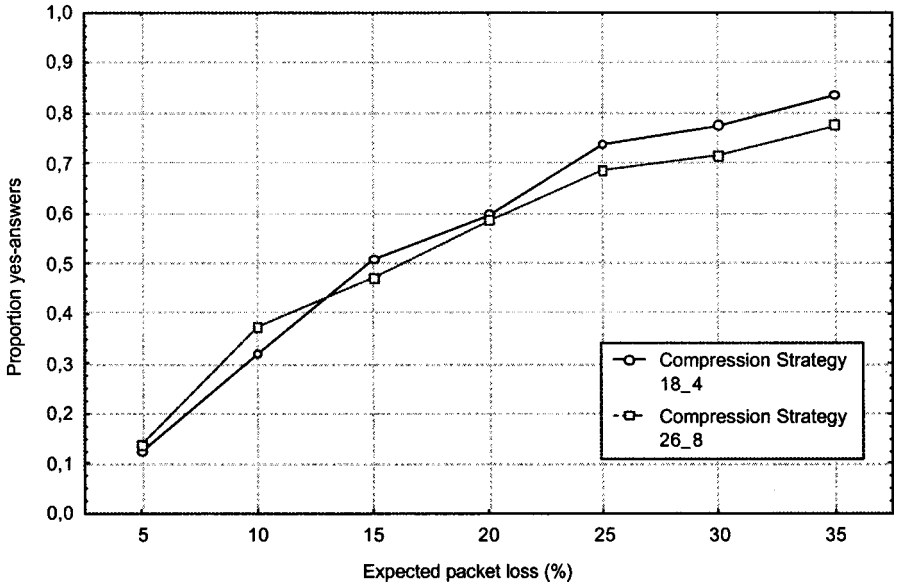

Fig. 6 Mean proportion of detection of distortion for the two compression strategies.

4 Results

4.1 *Analysis of Variance of Observers' Data*

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 of yes answers (see, e.g., Ref. 13, 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 participants are considered as the random factor. As a criterion of a significant effect for all effects 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 compression strategy (C) $F(1,9)=0.35$ was not at the significance level chosen. The interactions of scenes with packet loss, abbreviated as S×L, and that of scenes with compression strategy, abbreviated as S×C, were also significant, $F(30,270)=4.12$ and $F(5,45)=3.33$, respectively. The interaction of packet loss with compression strategy, L×C, as well as the third-order interaction of scenes with packet loss and compression strategy, S×L×C 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 $F(1,9,0.95)=5.12$, as suggested by Winer.[14] A similar analysis for the nontransformed values was also done, with the same effects being significant.

The similarities for the two compression strategies are shown in Fig. 6 for the nontransformed detection values. It is likely that the similarities for the two strategies point to similar underlying psychophysical processes for the two scenes presented at the two levels, since differing appearances would point to dissimilar processes.

The results show 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 compression strategy, L×C, i.e., the effects of the packet losses were different for the two compression strategies. The effects for scenes at different compression strategies, i.e., S×C, may explain the significant differences between the scenes, although the main effect of the two compression strategies, C, do not. As mentioned, another interaction effect with compression strategy, namely, that with packet loss, L×C, was also significant. These interaction effects mentioned here are illustrated in Fig. 7.

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. As a display model, the earlier measured gamma function of the monitor was used. The 50% detection threshold values for each participant and each image and compression strategy were determined by fitting a second-order polynomial to the data, and finding the packet loss corresponding to the 50% detection value. The data were not corrected for guessing. One reason was that a comparison between the thresholds computed with or without correction (see Ref. 12, p. 81) showed only minor differences. The average threshold for each scene was calculated only for those participants who had a 50% point falling within the bounds established by a second-order polynomial. The average thresholds in terms of packet loss for all the participants together with the predicted values given by the model are shown in Fig. 8.

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 second-degree polynomial was fitted between the model values in a loga-

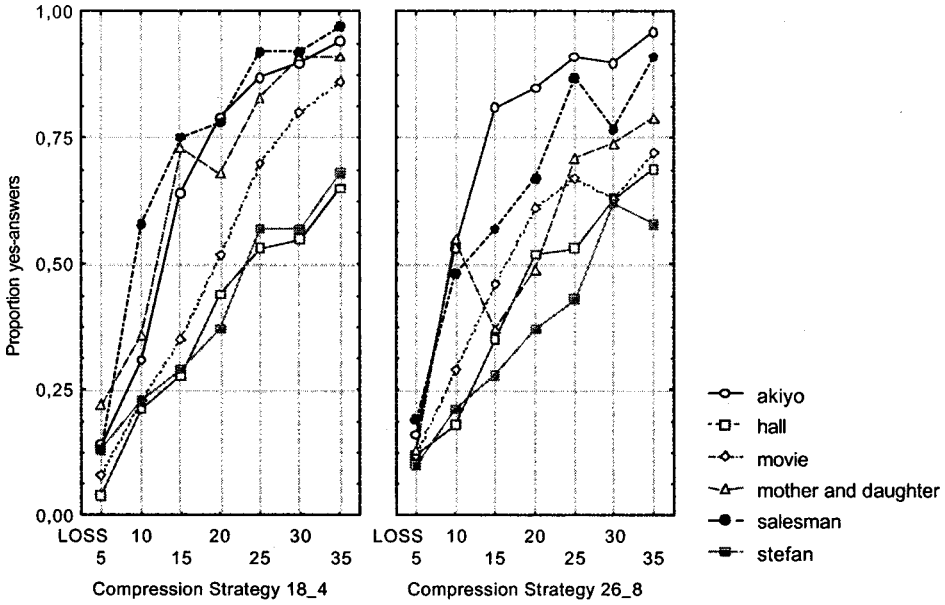

Fig. 7 Mean effects of packet losses for the two compression strategies for the six scenes.

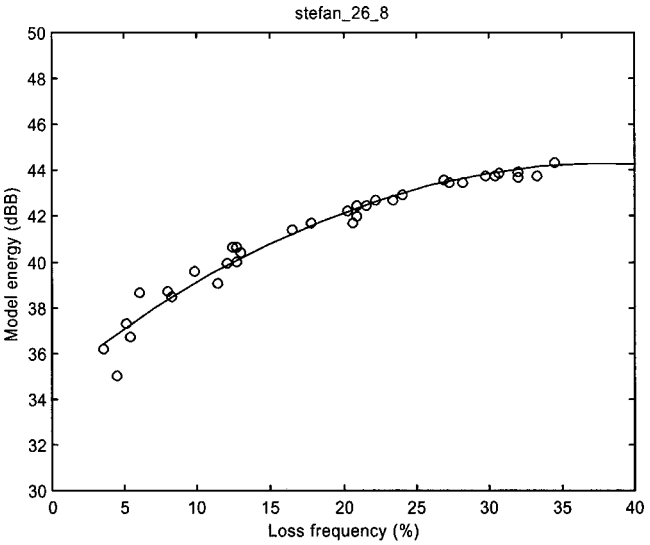

Fig. 8 Second-order polynomial fitted between the model predictions in logarithmic scale and the actual packet loss frequencies, which are used to estimate the packet loss frequency, which gives a just-noticeable difference from the reference level of 5% loss.

rithmic scale and the packet loss frequencies of the actual loss (see Fig. 9). The logarithmic scale used is called decibel Barlow (dBB) and is a scale of contrast energy, which is related to the weakest stimuli that a human can see.[15] The thresholds were then estimated as the packet loss difference from the 5% level, which gave a model difference of 1 jnd, i.e., just-notiecable difference. The correlation between these values of the model and the average empirical thresholds was 0.66 for compression strategy 18_4, 0.80 for compression strategy 26_8, and 0.56 for both compression strategies together. This illustrates a fair ability of the model to predict the differential effects of distortion for the different images.

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

The threshold values based on the PSNR values were calculated and the resulting values are also shown in Fig. 8. The correlation between the PSNR and the average thresholds was 0.45 for compression strategy 18_4, 0.62 for compression strategy 26_8, and 0.47 for both compression

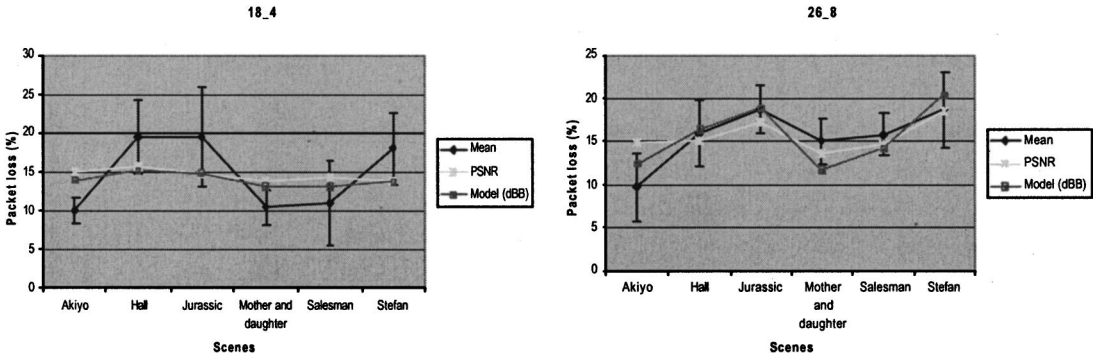

Fig. 9 Empirical mean 50% detection distortion thresholds for the subjects and model-based thresholds.

Table 1 Summary of multiple regression on the average yes answers for independent variables and energy or physically based measures.

	Beta*	*B*	*t* (410)	*p* level
Intercept		−55.7±10.70	−5.21	<0.001
Scene	0.29±0.08	0.10±0.03	3.61	<0.001
Compression strategy	0.16±0.06	0.19±0.07	2.56	0.011
Scene 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

*For the Beta and *B* values are also shown the standard errors.

strategies together, illustrating a slightly better ability of the visual model to predict the empirical values. The main difference in correlation between the model and PSNR comes primarily from the larger vertical 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 statistic.[13] The z value for the difference between the correlations for the model and PSNR for all the scenes 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 gives 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.[3] In the current calculation, the model output is in the form of difference contrast energy.

4.3 *Analysis of Version Dependence of Packet Loss*

The packet loss may fall in various places, due to chance. This is an uncontrolled variability on top of the intended packet loss settings. To minimize spurious effects of packet loss, the same scene with the same packet loss were made in different versions. The variable packet loss varied with a probability from 5 to 35%. To prevent accidental and strange instances of packet loss for a certain scene to dominate observer results, we had five different versions of each scene at each packet loss level and each compression strategy. 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. This is mostly a consequence of the short time sequences that were used. One way to get a measure of the actual extent of the distortion is to measure the contrast energy, by computing $E_c = At\Sigma_{x,y,t}c(x,y,t)^2$ rad^2 s, where A is the area of 1 pixel in degrees squared, and t is 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 $c(x,y,t) = L_d(x,y,t)/L_o(x,y,t)-1$, 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 version of every image presented. The average proportions of yes answers for the 10 participants then were computed for each version. The correlation between the dependent variable average proportion of yes answers and packet loss, contrast energy in decibel units, and contrast energy on a linear scale were 0.74, 0.26, and 0.24, respectively.

A multiple regression was performed with the number of yes answers of all the participants as the dependent variable and a number of variables (see Table 1), as the independent variables, including the PSNR values [see Eq. (1)]. The result is presented in Table 1, where B and beta are the nonstandardized and the standardized regression coefficients, respectively.

The R correlation was 0.86 and R^2, showing the explained variance, was 0.74. As we can see, 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 images presented when seen as a total. The actual packet loss and the PSNR measure are the most important parameters in this analysis.

To see how these measures compare to the model-based values, the correlation between the empirical values and the theoretical ones were computed for each scene and compression strategy. In addition, the correlations with the expected loss frequencies were also computed. Each value is based on 35 observations. The PSNR has a negative correlation since its value decreases with increasing distortion. The resulting correlations between the dependent variable and the various physical measures are shown in Table 2. This table shows that when the image sequence is held constant, the model-based values correlate slightly better with the participant's judgments than do the loss measures

[†]This computation is problematic if the luminance becomes zero, but the lowest luminance used in this experiment was 0.1 cd/m^2.

Table 2 Correlations between the mean of yes answers of the subjects to physical values of image distortion.

Scene	Compression Strategy	Expected Loss	Actual Loss	Contrast Energy (dBB)	Contrast Energy (linear)	Model (Ahumada)	PSNR (Absolute 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
"Movie"	18_4	0.93	0.93	0.89	0.86	0.93	0.92
"Movie"	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

(expected or actual loss) or the contrast energy (dBB) measures. However, the PSNR values appear to be on a par with the model values. These correlations are based on the whole range of the responses of the participants, whereas the threshold values look at only one point of the scale for the determination. Two curves may have a high correlation, while also being far apart. However, the threshold is only one value on the curve, and therefore the predictions of the model need only to be close to the empirical responses in the vicinity of the threshold. Furthermore, the correlation of the values for all the scenes with the predictions of the model tell us more about the success of the model than does the closeness of the prediction to a single scene.

4.4 *Category Scales*

The values for each of the four category scales were analyzed by analysis of variance, using a mixed model, i.e., a model with both random and fixed factors, with the participants seen as the random factor. The main factors in each analysis were participants, compression strategy, and scenes with degrees of freedom of 9, 1, and 5, respectively. The interaction of compression strategy with scenes thus had 5 degrees of freedom. There is only 1 degree of freedom for the compression strategy, since this value is equal to $p-1$, where p is the number of levels, in this case 2. Some of the F ratios of the analysis for the category scales are shown in Table 3. The participants factor and its interactions are not shown.

The average values of blockiness, noise, blur, and total impression for all participants for the scenes at the two compression strategies are shown in Fig. 10. We include here the judgments of the scenes as a group, although this judgment was not included in the analysis of variances. It is apparent that the "Mother and Daughter" scene was most sensitive to disturbances for both presentations. This is most evident from the category scale called 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 "Stefan" has a low total impression on both presentation levels. In general, the values of the 26_8 compression strategy are higher, i.e., a lower image quality, than the 18_4 compression strategy. One sees a difference in the values determined for the images, where the 26_8 generally had worse impressions than the 18_4 compression strategy. However, the difference between the two compression strategies was significant only for the blur scale. Note 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 participants ($n=102$ 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 and negative correlations as indicating that the two dependent variables refer to two

Table 3 The F ratios resulting from the analysis of variance of the category scales for three of the effects.

	Blockiness	Noise	Blur	Total Impression
Compression strategy, 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, $p<0.05$
Compression strategy × scene, df=5	0.33	0.43	1.10	0.70

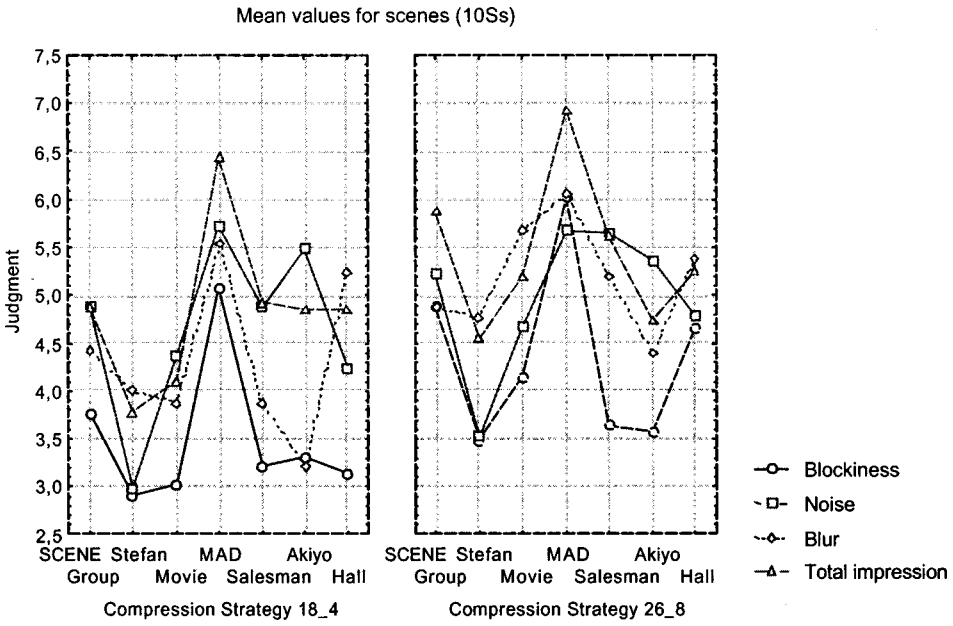

Fig. 10 Mean judgments of the subjects in experiment 2 for the scenes, as a group and separately, for the four category scales, at both compression strategies.

different psychological processes, one concerned with detection of distortions or disturbances, the other with overall quality judgments.

One can surmise that the most important parameter of the category scales in a practical situation is that for total impression. It tells, e.g., how satisfactory a user of a telecom service would be of its image quality. Seeing this scale as the dependent variable and performing a multiple linear regression on it with the other scales including that for the threshold values, would tell us about the predictive power of each of the dependent variables for total impression. This was done on the nonstandardized values for the 18_4 and 26_8 compression strategies. For the 18_4 compression strategy, the variables blockiness, noise, blur, and threshold ($n=53$) had coefficients 0.28, 0.48, 0.34, and 0.05 with the t values 3.63, 8.75, 4.80, and 2.81, respectively, all significant at $p<0.05$. For the 26_8 compression strategy, the variables blockiness, noise, blur, and threshold ($n=49$) had coefficients 0.41, 0.28, 0.40, and 0.03 with the t values 5.65, 3.86, 4.44, and 1.31, respectively. The three first values, but not that for the threshold, were significant at $p<0.05$.

The linear regressions indicate that the two compression strategies were judged differently. The most important variable for the 18_4 level was noise, while the most important for the 26_8 level was blockiness. The threshold value was not important for the judgment of total impression.

5 Discussion

For experiment 1 the spatiotemporal model was a good predictor of distortion detection thresholds for three of the video image sequences. As shown in Fig. 1, for these three images the predictions of the model were close to the empirical mean. In Fig. 9, showing results for experiment 2 involving packet losses, the model failed to predict which of the image sequences would be the most affected by a packet loss for the compression strategy giving a better quality level, i.e., 18_4. 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 indicated the probable existence of two psychological processes for the judgment of the sequences.

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 participants after the study indicated that they had been unable to use the reference images presented at the beginning of each session. To study whether or not the persons used an inner reference, one must 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'' has been judged as having the best overall quality as a reference image (see Fig. 10), and shows very low sensitivity for increasing rates of packet losses (see Fig. 7). This is also the scene that contains the most movements. Another important aspect is the sensitivity of the human visual system for disturbances located in faces. Attempts were 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 (see Ref. 16). From Fig. 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 image-based measures of the image sequences: contrast energy, spatiotemporal model energy, and PSNR. The PSNR values predicted the differences between the image sequences almost as well as the spatiotem-

poral model. This could be an effect of the frame rate of the second experiment, which was 15 Hz. The temporal contrast sensitivity function (CSF) does not influence the results that much, especially when considering the Nyquist frequency of about 8 Hz, which is near the peak of the temporal CSF. However, the spatial CSF should still have some influence, albeit limited, since the Nyquist frequency for the spatial domain was about 12 cycles/deg. The problem must be investigated further for an explanation. This result is consistent with the result of the Video Quality Expert Group[17] (VQEG), even though different visual models were used, some of the models were based on the same principle as the model investigated here.

One difference in the two experiments is that the three participants in the first study were all active vision researchers, whereas the participants in the second study had more varied backgrounds. Another important difference is that in experiment 1, the thresholds were 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. To be able to calculate the threshold because of the character of packet losses, we could not use the coded reference scenes with its packet losses, but the model looked at the difference between the coded original scene and the distorted (and coded) 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[18] that once distortions are above threshold, different persons can weigh different dimensions differently so that no single measure of image quality is possible. The category scaling in this study also illustrates this.

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, for example, not much different from the average of the other scenes regarding the detection values, but is much different on the category scales. We believe that this mirrors the existence of two psychological processes. One is used to detect 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 consider that it is of importance for the quality impression. Perhaps it is similar to the distinction between local and global psychophysics,[6] where the former refers to methods aimed at determining detection and discrimination thresholds, while the latter refers to what are called supra-threshold measurements. Roufs and Boschman[19] point out 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 somewhat close to the empirical values for some of the conditions. 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 compared to simpler measures such as the PSNR ought to be further investigated.

Acknowledgments

We thank Al Ahumada, National Aeronautics and Space Administration (NASA) Ames Research Center, for help and discussions on the experiment and on the model used. We are grateful to advice provided us by Sture Eriksson, Uppsala University, and by Birgitta Berglund, Stockholm University. We thank Jean Bennett, visiting professor at ACREO, for helpful advice. Börje Andrén was helpful, especially in conducting photometric measurements. The observers' contributions are gratefully acknowledged. Anonymous reviewers to earlier versions of this manuscript gave valuable comments. This work was supported by the companies Telia Research AB, Ericsson SAAB Avionics AB (now renamed SAAB Avionics AB), Celsius Tech Electronics AB (now renamed FLIR Systems AB), FMV, and by NUTEK (Swedish Board for Industrial and Technical Development, now renamed VINNOVA).

References

1. A. B. Watson, J. Hu, J. F. McGowan III, and J. B. Mulligan, ''Design and performance of a digital video quality metric,'' in *Human Vision and Electronic Imaging IV*, B. Rogowitz and T. N. Pappas, Eds., *Proc. SPIE* **3644**, 168–174 (1999).
2. S. Winkler, ''Perceptual distortion metric for digital color video,'' in *Human Vision and Electronic Imaging IV*, B. Rogowitz and T. N. Pappas, Eds., *Proc. SPIE* **3644**, 175–184 (1999).
3. A. J. Ahumada, Jr., B. L. Beard, and R. Eriksson. ''Spatio-temporal image discrimination model predicts temporal masking functions,'' in *Human Vision and Electronic Imaging III*, B. Rogowitz and T. N. Pappas, Eds., *Proc. SPIE* **3299**, 120–127 (1998).
4. K. Brunnström, R. Eriksson, B. Schenkman, and B. Andrén, ''Comparison of predictions of a spatiotemporal model with responses of observers for moving images,'' Technical Report, TR-338, Institute Optical Research, Kista, Sweden (1999).
5. G. Borg, ''Psychophysical judgment and the process of perception,'' in *Psychophysical Judgment and the Process of Perception*, H.-G. Geissler and P. Petzold, Eds., VEB Deutscher Verlag der Wissenschaft, Berlin (1982).
6. J.-B. Martens and M. Boschman, ''The psychophysical measurement of image quality,'' in *Vision Models and Applications to Image and Video Processing*, C. J. van den Branden Lambrecht, Ed., 71–101, Kluwer, Boston, MA (2001).
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. D. Pearson, ''Viewer response of time-varying video quality,'' in *Human Vision and Electronic Imaging III*, B. Rogowitz and T. N. Pappas, Eds., *Proc. SPIE* **3299**, 16–25 (1998).
11. H. de Ridder and R. Hamberg, ''Continuous assessment of image quality,'' *SMPTE J.* **106**(2), 123–128 (1997).
12. G. A. Gescheider, *Psychophysics: Method, Theory and Application*, Lawrence Erlbaum, Hillsdale, NJ (1985).
13. D. C. Howell, *Statistical Methods for Psychology*, Duxbury Press, Belmont, CA (1997).
14. B. J. Winer, *Statistical Principles in Experimental Design*, McGraw-Hill, New York, (1962).
15. A. B. Watson, H. B. Barlow, and J. G. Robson, ''What does the eye see best?'' *Nature (London)* **302**(5907), 419–422 (1983).
16. S. Daly, K. Matthews, and J. Ribas-Corbera, ''Visual eccentricity models in face-based video-compression,'' in *Human Vision and Electronic Imaging IV*, B. Rogowitz and T. N. Pappas, Eds., *Proc. SPIE* **3644**, 152–166 (1999).
17. VQEG, ''Final report from the video quality experts group on the validation of objective models of video quality assessment,'' March 2000, URL: http://www.crc.ca/vqeg.
18. A. J. Ahumada, Jr. and C. Null, ''Image quality: a multidimensional problem,'' in *Digital Images and Human Vision*, A. Watson, Ed., MIT Press, Cambridge, MA (1993).

19. J. A. J. Roufs and M. C. Boschman, ''Visual comfort and performance,'' in *The Man-Machine Interface*, J. A. J. Roufs, Ed., Macmillan, London (1991).

Kjell Brunnström received his MS degree in engineering physics and his PhD degree in computer science from the Royal Institute of Technology, Stockholm, Sweden, in 1984 and 1993, respectively. From October 1985 to April 1987 he was a visiting research student with Tokyo University, Japan. During 1995 he was a postdoctoral associate with the University of Surrey, Guildford, United Kingdom. He is currently a researcher with the research institute ACREO, previously called the Institute of Optical Research, Stockholm. His current main research interest is image quality assessment for still images and video.

Bo N. Schenkman received his BA degree in psychology and philosophy from the Hebrew University, Jerusalem, Israel, and his PhD degree in psychology from Uppsala University, Sweden, in 1985. From 1985 to 1996 he was a human factors specialist in the R&D Departments of the Swedish computer divisions of Ericsson, Nokia, and ICL. During 1996 he did research at the Royal Institute of Technology, Stockholm, on image quality issues. From 1997 to 1998 he worked with telecommunication research at Telia, Stockholm. In 1999 he joined the Institute of Optical Research in Stockholm, later named ACREO. His present research interests are human perception, image quality, human factors, and psychophysics.