Off-line evaluations of noise reduction algorithms for fluoroscopy

Institute for Perception Research
PO Box 513, 5600 MB Eindhoven

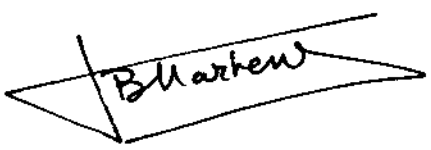
04.01.1996

Rapport no. 1087

Off-line evaluations of noise reduction algorithms for fluoroscopy

W.M.C.J. van Overveld

Voor akkoord: Dr.ir. J.B.O.S. Martens



Off-line evaluations of noise reduction algorithms for fluoroscopy

W.M.C.J. van Overveld

Index

Summary

This report describes the evaluations of noise reduction algorithms carried out in the context of the IPO/PMSN project "Optimization of the perceived quality of fluoroscopic images". Four new algorithms have been applied to two fluoroscopic image sequences and the effects have been evaluated and compared to the existing processing using paired comparisons. Multi-dimensional scaling analysis provides a lucid way to compare the various algorithms w.r.t. perceived noisiness, sharpness, contrast, and overall quality of the sequences processed with these algorithms. The main results of the evaluations are the following.

From the pairwise comparison experiment, it turns out that the recursive temporal filter performs best, closely followed by the 3-d order statistics filter. These two algorithms reduce a large part of the noise while maintaining a fair sharpness and contrast (although less than the original, unprocessed sequence). However, the observers' remarks indicate that the noise breakthrough artefacts in these algorithms may be extremely objectionable. The other two (purely spatial) algorithms reduce less noise. The multiresolution FMH algorithm appears to introduce unsharpness and contrast loss, and the spectral estimation algorithm has different perceptual effects depending on the amount of noise in the original sequence.

The results from the pairwise comparisons can be adequately described in a two-dimensional perceptual space. An experiment on dissimilarity ratings was used to transform this space to a geometrically meaningful space. In general, the experiments show that the quality impression is strongly correlated with the impression of "absence of noise". Similarly, contrast and sharpness impressions are closely related. Still there are marked individual differences; not only with respect to quality (personal taste) but also for the contrast and sharpness results. Both the differences and the similarities between the results of individual subjects are discussed in this report.

1 Introduction

This document is a deliverable of the research project "Optimization of the perceived quality of fluoroscopic images". It describes the results of a first evaluation of various noise reduction algorithms, which have been developed in the accompanying "Image processing for fluoroscopy" project.

The main aim of these evaluations was to get some insight in the effect different kinds of algorithms have on image quality, so that further work may be concentrated on the most promising algorithms and possible flaws of the algorithms might be corrected. A second goal was to find out how different aspects of image quality like perceived sharpness, noisiness and contrast affect the overall quality judgement of fluoroscopic image sequences treated with noise reduction algorithms. We concentrated on the cosmetic effects of the algorithms rather than the diagnostic (performance-oriented) effects, for reasons which are mentioned in Section 2. In Section 2 we also explain how the experiments were set up.

Section 3 focuses on the data analysis. Here we describe how the raw data (preferences for either the left or the right image in a split screen display) are transformed into quality, noisiness, sharpness and contrast scores on an interval scale, and how these scores are subsequently used to map the stimuli (i.e., the effect of each algorithm on each sequence) into a two-dimensional perceptual space.

In Section 4, we discuss the remarks made by the subjects and we relate these to the results found in Section 3. From this we draw our final conclusions. Furthermore we mention some consequences for future evaluations and for further research into noise reduction algorithms.

2 Experimental set-up

2.1 Algorithms

As said in the introduction, at this phase of the project our main interest is to select those noise reduction techniques which are worthwhile for further investigation and/or hardware implementation. We only need to have a general ideal of the usefulness of the algorithms; it is not necessary to find the optimum setting of parameters for each of the algorithms, because there is still time for that at a later stage (after hardware implementation). Note that this implies that the algorithms evaluated here may not have been used in the optimum setting, so that the comparison is not completely fair. Although we are aware of this restriction, we do not think it is a severe drawback for the goal we wanted to achieve.

We have evaluated the following four algorithms, all of which have shown to be promising in earlier informal evaluations. The algorithms also have in common that it is feasible to implement them in hardware (although some of them are more complex than others).

- The Multiresolution Conditional FMH (FIR-Median Hybrid) filter, abbreviated by MR-CFMH. This is a spatial linear/median hybrid filter which is applied on different subbands of an image using a Laplacian pyramid decomposition. The algorithm is fully described in [AK95b]. The version which was evaluated has the following details. The pyramid consisted of a four-level-decomposition (three filtered levels). A simple 3x3 separable binomial kernel was used for decimation and interpolation. At all levels, two iterations of the 5x5 spatially recursive FMH were used, except for the finest level in the case of sequence n431 (see Section 2.3): here only one iteration was used. The threshold factor (to be multiplied by the standard deviation) was 2.5. Furthermore, for the sequence xtv8_001, a slight enhancement was activated during reconstruction: the two finest levels were multiplied by 1.1.
- The spectral estimation algorithm. In this algorithm, a Fourier (or alternatively, a DCT) transformation is applied to overlapping blocks of an image. The FFT coefficients are attenuated according to the magnitude of the coefficient and the estimated noise power spectrum at the given intensity and the given spatial frequency: the so-called "noise curve". The algorithm is further explained in [AK95a]. We evaluated the version with the FFT transform, for block size 64x64 with 16 pixel overlap. No enhancement or explicit directional sensitivity were used. The noise curve used is the sigmoid curve described in [AK95a], equation (3.13) applied with $\alpha = 3.3$.
- A recursive temporal filter similar to the one which will be implemented in the DSI-5 system and which is described in [Flo95]. The filter is a temporal anisotropic diffusion filter, combined with a so-called gate image which is an FMH filtered version of the input image at time t. The gate image is compared with the filtered frame at time t-1 and t+1 to determine the "probability" that a discontinuity has occurred. This probability is also based on the expected noise level, which is derived from the intensity in the gate image through a look-up table. The Kalman-like gain factors k_c (causal) and k_a (anti-causal) at time t, which determine how much the output frames at time t-1 and t+1, respectively, contribute to the output at time t, are computed from the causal gain factor at time t-1 and from the discontinuity probability. Discontinuity is detected by taking

the normalized difference between the current gate image and the previous filtered frame. No motion estimation was implemented. The difference with the DSI-5 algorithm is the inclusion of the "future frame" (at time t+1), and the fact that the current point's weight can be modulated depending on noise peak detection. Parameter values used are: $\alpha_{max} = \gamma_{max} = 0.7$ (the causal and anticausal integration factors), $\beta_{min} = 0.1$ (the modulation factor for the current point) and t = 2.0 (the cut-off point of the discontinuity curves).

- A 3-d order statistic filter, as described in the minutes of the kernel team meeting of the project "Image processing for fluoroscopy" at October 26, 1995 (ref. XDB 048.95.0258 FS/FS). This is the least complex algorithm, since it is composed of 2-d (spatial dimensions) and 1-d (temporal dimension) linear filters only. The filter uses three temporally consecutive frames which are first preprocessed by an FMH filter (2LH2D, w=1). The filtered versions are denoted by g_{t-1}, g_t and g_{t+1}. The output of the filter is equal to g_t if both $|g_t - g_{t-1}| / sig(t) > T$ and $|g_t - g_{t+1}| / sig(t) > T$, where sig(t) is a normalization coefficient and the threshold T equals 1.5. If either difference is smaller than T, the output is set to the median of g_{t-1}, g_t and g_{t+1}.

In our evaluations, we also included an extra "algorithm" for reference purposes, namely the processing that currently takes place on fluoroscopy images. In other words, an original image sequence without further processing will be treated as a sequence processed with "algorithm 1". The new algorithms to be tested are referred to as numbers 2, 3, 4 and 5, according to Table 1.

no.	algorithm
1	reference
2	MR-CFMH
3	spectral estimation
4	temporal
5	3-d order statistics

TABLE 1. The noise reduction algorithms to be evaluated. See text for details.

2.2 Perceptual attributes to be evaluated

Obviously, perceived image quality is of the main interest if we want to decide with which algorithms the project should be continued. However, we know (cf. [Kun86], [Rou92], [Ove95b]) that there are two types of image quality: appreciation-oriented (cosmetic) quality and performance-oriented (diagnostic) quality. We have chosen to evaluate cosmetic quality only, for reasons stated below.

Informal assessment of the effect of the various algorithms has lead to the belief that the algorithms will largely affect the cosmetic aspects of image quality, and the performance-oriented quality (the ability to see subtle anatomical or pathological details) will be affected much less.[1] For this reason, and also because we only need (and can only get) a "rough" idea of the possible success of the algorithms, we decided to evaluate appreciation-oriented image quality only. Additional advantages are:

- the generation of stimuli is much more complex if subjects have to perform a task with the stimuli. This would e.g. involve selecting or adding subtle details to be detected by subjects, similar to the experimental set-up used in [Ove95a].
- experimental sessions will take longer if the subjects have to search for some sub-threshold detail than if they are asked about a "first impression" of the image quality.
- if we do not use a true clinical task, it is not necessary to get the cooperation of radiologists (although performance can be measured with non-experts as well (cf. [Ove95a])).

Apart from quality, we are also interested in the "components" of the quality. If a subject calls a sequence "bad", is that because it contains too much noise (not enough noise reduction), or because it is flat or blurred (too rigorous noise reduction)? This is interesting from a fundamental point of view - studying how a general quality impression can be described as a function of underlying perceptual attributes - but the extra information can also help to suggest the direction in which the algorithms may be improved.

In general, sharpness, noisiness and contrast are important components of image quality (cf. [MN75], [ACSK90]) and it is also reasonable to expect that noise reduction algorithms affect these perceptual attributes. For that reason we ask the viewers' opinion about the perceived sharpness, noisiness and contrast of the sequences. Other factors may play a role as well (see [Ove95b] for a whole range of factors that may affect the quality of fluoroscopy sequences) but we have to make a compromise. We will assume that any other attributes (e.g. "patchiness") are secondary effects; it is just not possible to evaluate too many attributes in one experiment. Anyway, we can check whether the quality scores can really be described as a function of these parameters, through the use of principal component analysis (Section 3.2). Apart from this, the subjects could freely comment on any aspect of the sequences shown to them in the experiment. Their remarks also indicated to which amount quality was really determined by noise, sharpness and contrast.

1. Performance may be affected when a physician has to look at noisy or blurred images, even if all details are still visible, when it is fatiguing to look at such images for a prolonged time. Such aspects are however taken into account when assessing cosmetic quality as was done in this experiment.

2.3 Image material

To produce stimuli, the five algorithms have been applied to two different image sequences, hereafter called "scenes". The scenes were chosen to represent typical medical applications with certain "difficult" aspects like fast local and global motion, small low-contrast details and a high noise level.

We chose sequence xtv8_001 because it is the most noisy sequence of all sequences in the image data base and thus poses a real challenge to the noise reduction algorithms. A drawback of this scene is that the source of the noise is a little dubious: some external noise may be included as a result of the recording of the sequence. The scene has interesting features like a moving catheter and a contrast injection. It also has a fairly homogeneous distribution of grey levels.

The second scene chosen was n431. This sequence from a colon examination was chosen because it also has a fair amount of noise (about half of the noise in the first sequence, in terms of noise variance). The overall contrast in this sequence is higher than in xtv8_001, and the type of features is different: the thin folds in the colon wall (in the "double contrast" part of the image) form high frequency, high contrast details. There are also very dense, black regions in this image: the part where the bowels are still filled with barium ("single contrast"). This scene has some very fast table motion. As seen in Table 2, the scenes are referred to as scene 1 and scene 2. Scene 1 consists of 213 frames and scene 2 has 47 frames. Each frame consists of 512 x 512 pixels of 8 bit grey value. Appendix D shows representative frames from each of the two (unprocessed) sequences.

no.	scene
1	xtv8_001
2	n431

TABLE 2. The scenes used in the evaluations. See text for details.

Note that we have only considered continuous fluoroscopy sequences, recorded with an xtv8 camera. One of the reasons for this is that all other sequences available at this moment are much less noisy, so that all algorithms can reduce the noise in those sequences without great difficulty. Thus it is hard to discriminate between the perceptual effects of different algorithms when they are applied to such "easy" sequences. Different types of image material (different camera, higher spatial resolution, pulsed fluoroscopy at various frame rates, ...) should be considered at a later stage, e.g. when fine-tuning of the parameters for a few isolated algorithms is called for.

2.4 Method

All frames of the two sequences were processed by each of the algorithms in Table 1. Each combination of a scene and an algorithm formed one stimulus in the experiment. Thus we had 2 x 5 = 10 stimuli. The stimuli were presented in pairs, side by side in a split screen. Thus only half of each image was displayed: a 512 x 256 vertically oriented subimage which was considered to be a sufficiently "critical" and representative part of the whole 512 x 512 image. For the xtv8_001 scene, we chose the subimage having the pixel with coordinates (190,1) as the upper left-hand corner. This ensured that the border of the shutter was still visible, which was important because certain artefacts might show up at this high contrast border. For the n431 scene, we used pixel (120,1) as the upper left-hand corner of the subimage, which approximately corresponded to the centre part of the image.

The two stimuli in each pair were derived from the same scene and only varied in the algorithm applied to them. Stimuli were displayed using the ISP display system of PMSN and the command language "Divise" (cf. [SL93]). The low-pass filter available in this system was switched off. The stimuli in a pair were synchronized and shown in repeat mode (frames 1, 2, ..., n, 1, 2, ..., n and so on). Shuttle mode (frames 1, 2, ..., n, n, n-1, ..., 1) would have produced a smoother transition from one set of n frames to the next. This would have been more comfortable to look at, especially for the short sequence n431, but effects of temporal filtering (e.g. "noise tails") could not be judged reliably that way. While a stimulus pair was being displayed, the observer answered four questions:

- which sequence, left or right, do you prefer in general?
- which sequence, left or right, is the least noisy?
- which sequence, left or right, is the sharpest?
- which sequence, left or right, has the most contrast?

Each pair remained on the screen for as long as it took for the observer to answer the questions for that pair; on average, this took about 15 seconds. All in all, 40 different pairs of stimuli were displayed: for each of the two scenes, every ordered combination of two different algorithms occurred. Thus if a pair with algorithm i on the left side of the screen and algorithm j on the right occurred, then the pair with algorithm j on the left and i on the right occurred as well. This was done to preclude possible biases of subjects for one half of the screen, and to correct for inhomogeneities in the screen. Each of the 40 pairs was displayed twice, so that every pair of stimuli was compared 4 times by each subject. A session took between 30 and 45 minutes, including overhead time for instructions and for loading and synchronizing sequences. The stimulus pairs of a single scene were displayed in pseudo-random order, balanced according to the rule given in [DR83] and counterbalanced for temporal order effects (i.e. if pair (a,b) was directly followed by (c,d) in one part of the session, (c,d) was also followed by (a,b) in another part of the session.). The stimulus pairs corresponding to scene 1 and scene 2 were interleaved within one session.

The reasons for presenting the stimuli in pairs and asking for a "left" or "right" answer, instead of e.g. presenting them one by one and asking for ratings on a scale from 1 to 10 (as was done previously: cf. [Ove93a], [Ove95a]) are as follows. First of all, in the earlier

studies some subjects - not used to scaling experiments - had difficulties in expressing their quality impressions in numbers. It was "unnatural" for them to use the whole range (1 to 10) when the perceived differences between stimuli were small. They also hesitated to assign a 10 to a stimulus which was less than "perfect", even though it was the best one (or the "least bad" one) among all stimuli presented. Thus we felt that asking for a "left" or "right" response would make things easier for the subjects.

The two-alternative forced choice (2AFC) paradigm used here is also a more direct measurement of the observer's sensation, since we circumvent the unknown mapping from the "sensorial strength" of an attribute to the numerical response given by the user. Of course, we will have to make assumptions about such a mapping later on if we want to analyse the data at anything higher than an ordinal level, but these assumptions are more explicit and can be verified. We come back to this issue in Section 3.1.

Thirdly, the 2AFC method is also more sensitive to small differences between stimuli. We are inevitably dealing with small differences in sharpness and contrast in our stimuli, since the goal of the noise reduction algorithms is in fact to leave the sharpness and contrast unchanged and only affect the noisiness. With stimuli presented one by one, the differences in contrast and sharpness (and in some cases also the differences in noisiness) would be very hard to notice, so that the ratings would all be the same in a scaling experiment. Even when the stimuli are presented side by side, subjects could still be inclined to assign equal scores to them because the differences are not visible at a first glance. Only when they are forced to select one of the two images, the small near-threshold differences in contrast or sharpness could tip the balance in favour of one of them.

After all stimulus pairs were displayed in the 2AFC experiment and the observer had responded to them, the ten stimuli were displayed separately, one at a time. The observer could make any remarks on the quality of each stimulus while it was being displayed; e.g. about the visibility of artefacts or about specific details he or she had been paying attention to. These remarks were used to see whether they could explain the preference choices in the paired comparison experiment. The remarks also contained valuable information for the possible improvement of the algorithms.

The location was the viewing room at PMSN, QJ-2. The monitor used was similar to monitors that are used in the clinical practice. The brightness and contrast settings on the monitor were adjusted such that most image details were visible both in bright and dark regions. The characteristics of the monitor after this adjustment, in terms of the grey-value-to-luminance curve, are given in Appendix C. The light in the room was dimmed. More specifically, the room was lit by some spotlights, which did not directly shine on the screen. We measured an illuminance of 82 lux on a part of the wall that was directly lit by one of the spotlights, and 3.3 lux was measured on the screen.

We used nine application specialists as subjects. The author of this report served as a tenth subject. The experiment was done by one observer at a time, so that they could not influence each other. The viewing distance was about 80 cm. The size of a pair of images on the screen was 26 cm (height) x 28.3 cm (width). For the given viewing distance, this amounts to a viewing angle of approximately 19 degrees.

3 Data analysis

3.1 From pairwise comparison to interval scaling

The data we collected in the 2AFC experiment described in the previous chapter can be written as sets of matrices $Q_s(i,j)$, $N_s(i,j)$, $S_s(i,j)$ and $C_s(i,j)$ where s ranges over the scenes (2 values) and i and j each range over the algorithms tested (5 values). The matrix Q_s contains the "general preference" or "quality" results for scene s, N_s contains the results for noisiness judged on scene s, S_s contains the sharpness results and C_s the contrast results. More specifically, for each s, i and j:

$Q_s(i,j)$ = the number of times scene s processed with algorithm j was *generally preferred* to the same scene processed with algorithm i in a paired comparison of these two stimuli;

$N_s(i,j)$ = the number of times scene s processed with algorithm j was judged *less noisy* than the same scene processed with algorithm i;

$S_s(i,j)$ = the number of times scene s processed with algorithm j was judged *sharper* than that scene processed with algorithm i;

$C_s(i,j)$ = the number of times scene s processed with algorithm j was judged to have *more contrast* than that scene processed with algorithm i.

The matrices can be considered for each subject separately, or they can be pooled over all subjects (see below). For each matrix, the ranking results from the paired comparisons can be transformed to an interval scale using three different methods: the "variance stable rank" method, the "Thurstone case V" method (both of which are described in e.g. [DR83]) and the "maximum likelihood" (ML) method (cf. [Tre68], Section 2.4).

The variance stable rank method is very easy to use and does not require a large amount of data, but it relies on the rather strict assumption that the number of times a stimulus is preferred to any other is proportional to the scale value of that stimulus. The fact that this method can be applied on small amounts of data enabled us to perform this transformation for the data of each subject separately. The Thurstone method, on the other hand, only assumes that reactions to a stimulus are normally distributed and derives scale values from the probability of confusing the order of two stimuli. This method was used in [Ove93a]. It requires a larger amount of observations than the variance stable rank method, to be able to estimate the parameters of the normal distribution from the observed frequencies. For this reason we only applied this method after pooling the results over all subjects. Another drawback of the method is the fact that it is not able to cope with a stimulus that is always judged lower than any other one. For our data, this occurred in the judgements of noisiness, where the original scene (algorithm 1) was always judged more noisy than any of the noise reduced scenes (algorithms 2 to 5). The maximum likelihood method uses the same assumptions as the Thurstone method, and has the same drawbacks, but the advantage of the ML method is that it also produces confidence regions: an indication for the accuracy of the estimated scale values. Again, this method was only applied to the data pooled over all subjects. The three methods are further explained in appendix A.

We have applied all three methods to our quality, sharpness and contrast data. The noisiness data was only treated with the variance stable rank method, for reasons explained above. All methods produced scores on interval scales, but we found separate scales for separate scenes, because different scenes were never directly compared in the experiment. Thus we obtained scores $Q_s(i)$, $N_s(i)$, $S_s(i)$ and $C_s(i)$ for every scene s and algorithm i (and, in the case of the variance stable rank method, for every subject). It should be noted that a high score on the Q_s, S_s or C_s scale means high general preference, high sharpness or high contrast, respectively, but a high score on the N_s scale means *low* noisiness. Thus high scores on all scales indicate aspects of high quality. Throughout this report, all scales were linearly transformed to a range of 1 to 10.

We first present the results of the separate subjects, as obtained by the variance stable rank method. It should be kept in mind that this is based on just four comparisons per pair of stimuli per subject, so that the reliability of these results should not be overestimated. Still it gives a fair indication on the level of agreement between subjects. The results are shown in Figure 1. It is obvious from this figure that the agreement on the noise impression is excellent, but it is less good on sharpness and even less on contrast. This is partly due to the fact that people judged these attributes based on different parts of the image sequences. For instance, for the n431 sequence it makes a difference whether sharpness is judged when the image is stationary or when it is moving: noise reduction algorithms can affect these two stages of the sequence in different ways. People also judged the perceived contrast in different ways depending on the image details they were looking at. Some mainly looked at the rendition of the "blackest black" in a barium sequence while others looked at intermediate grey tones e.g. in the interior of the double contrast colon or in rims of vertebrae. See Section 4 for a further discussion of this.

As was to be expected, preference judgements also vary between subjects. For instance, subject RC (second from the left in each cluster of bars) sometimes preferred the original image over the noise reduced ones, because he objected to the artificial look of the processed images. Other subjects did not mind the slight blur or lower contrast of the noise reduced images as long as they could still see relevant details. Such taste differences were found previously in informal assessments of static X-ray images (cf. [Ove93b]).

A closer look into individual differences can be obtained by clustering the results of subjects. We do this by first computing all correlation coefficients of the results of pairs of subjects (here we consider the "raw" results of a subject prior to the transformation to an interval scale: thus, the numbers in the lower left-hand corner of the frequency matrices $Q_s(i,j)$, $N_s(i,j)$, et cetera). We then combine the two subjects having the highest correlation coefficient, say subject 1 and 2, and consider them as one "cluster". We compute new correlation coefficients of this cluster (which is treated as a new "subject") with each of the other subjects s by just averaging the old correlation coefficients of subjects 1 and s on the one hand and 2 and s on the other hand. Again we look for the highest correlation coefficient and cluster the corresponding subjects. This procedure is repeated until all subjects end up in one cluster. We thus find a tree of clusters, where the highest nodes in the tree (closest to the leaves) correspond to the highest correlation.

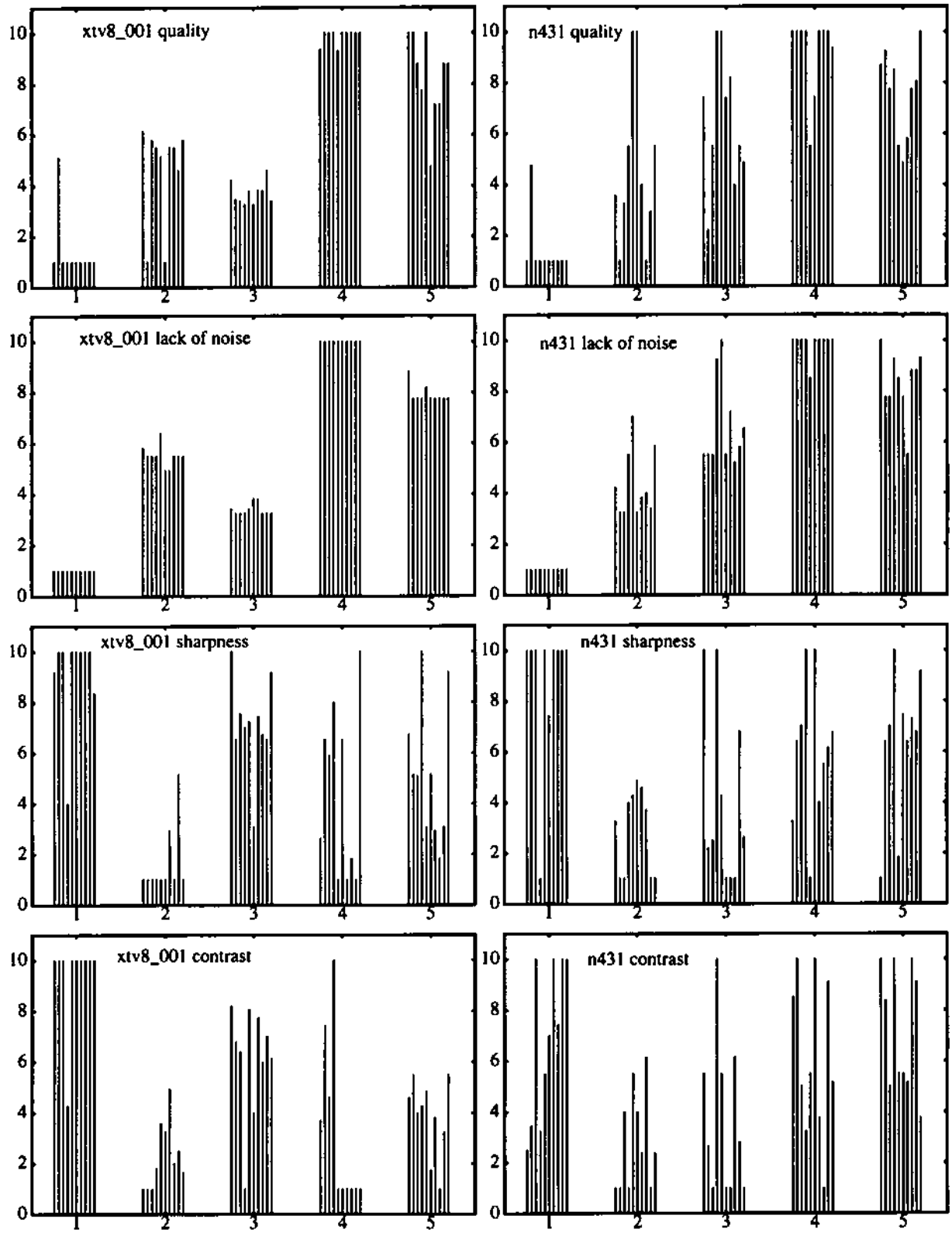


FIGURE 1. Individual scores on a scale from 1 to 10 for the ten subjects, plotted for each attribute and scene. The x-axis is indexed with the five algorithms. The initials of the subjects are, in the order in which their scores are shown from left to right: LV, RC, AB, RK, WC, MK, ER, PG, TL and IO.

These correlation trees are shown in Figure 2. An "overall" tree is shown in which correlation coefficients are computed using the complete data set of each subject. We also show trees based on the data for separate attributes; thus, using only correlations between the data in the $Q(i,j)$ matrices (pooled over the scenes), or the $N(i,j)$ matrices, et cetera.

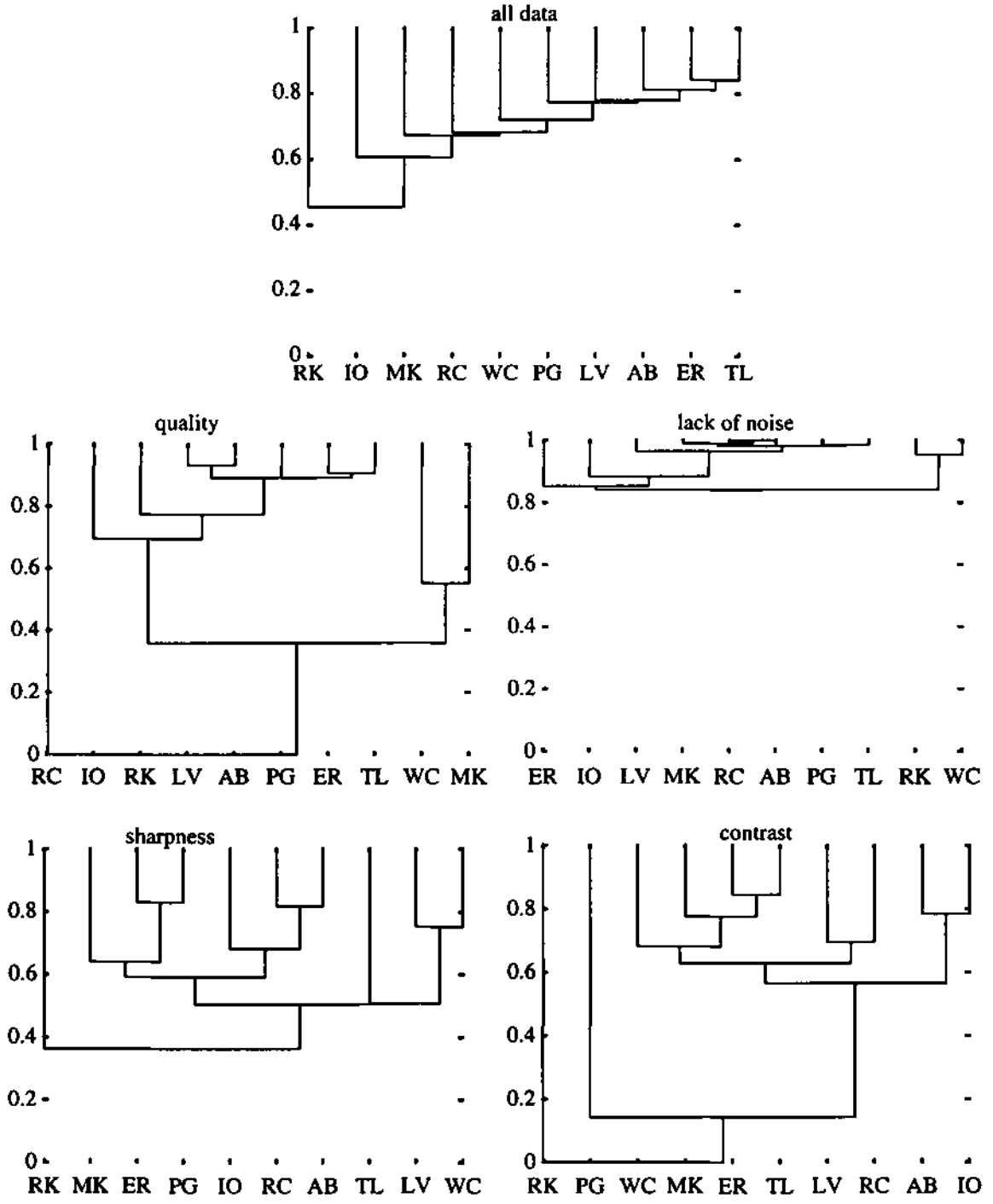


FIGURE 2. Clustering of subjects based on correlation. An "overall" clustering based on the complete data set is shown, as well as separate clusterings for the different attributes. The horizontal axis is labelled with the initials of the ten subjects; the vertical axis shows the correlation coefficient.

In the remainder of this section, we consider the results after pooling over the subjects. First we study the effect of the transformation method: variance stable rank, Thurstone case V or the maximum likelihood method. The effects are shown in Figure 3.

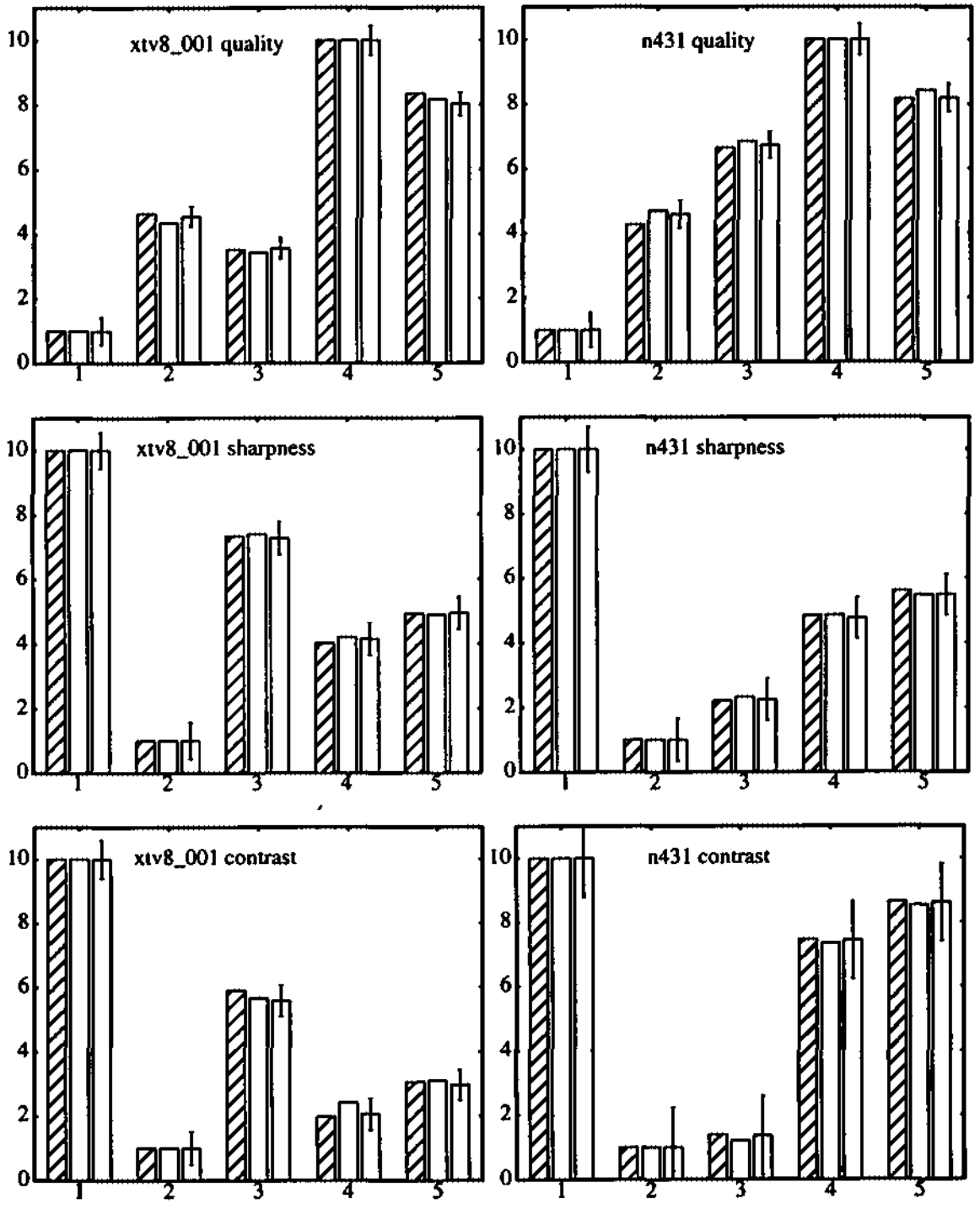


FIGURE 3. Scores derived with different transformation methods. Variance stable rank: hatched bars. Thurstone case V: dotted bars. Maximum likelihood: blank bars. The length of the error bars is twice the estimated standard deviation. Results are shown for different algorithms (indicated on the x-axis), scenes and attributes (shown in different bar charts).

We can only compare the results for quality, sharpness and contrast, because the Thurstone and ML methods could not be used for the noise data, as mentioned before. Still, the excellent correspondence of the results for the other attributes leads us to believe that the assumptions of the variance stable rank method are valid in our case. We will assume that this also holds for the noise data, so that we can safely proceed with the results of the variance stable rank method alone.

Figure 4 below shows the same results of the variance stable rank method pooled over subjects, but this time including the results for noisiness. In this figure, the effect of the scene contents is highlighted by comparing the results for different scenes in one graph. We point out, however, that scores should not be directly compared between scenes. A quality score of 3.8 for algorithm 3 applied to scene xtv8_001 and a score of 6.6 for the same algorithm applied to scene n431 does not imply that the algorithm performs "better" on xtv8_001 than on n431. We can only interpret the quality scores of algorithm 2 and 3, for instance, as: algorithm 2 performs somewhat better than algorithm 3 for scene xtv8_001, but the opposite is true for scene n431.

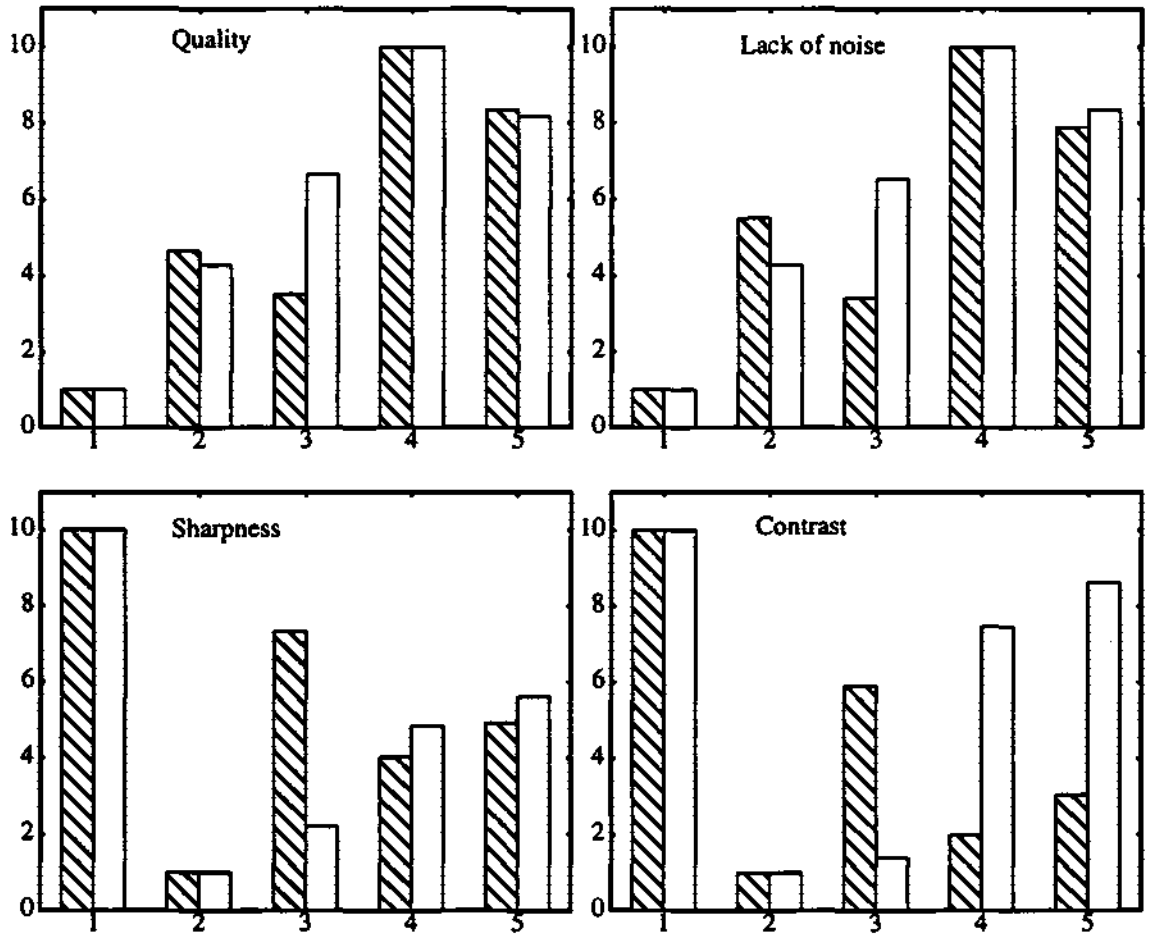


FIGURE 4. Attribute scores, as derived with the variance stable rank method, per algorithm. Scores are compared for the two scenes: xtv8_001 (hatched bars) and n431 (dotted bars).

Several conclusions can be drawn from Figure 4. Firstly, it is seen that quality and lack-of-noise scores are close to each other in every case. It is seen that algorithms 4 and 5 perform well on both scenes, in terms of the amount of noise reduction and the perceived quality. Algorithm 2 does not perform very well on either scene. The effects of algorithm 3 with respect to quality and noisiness are different for the two scenes: for these two attributes, algorithm 3 is ranked higher among the algorithms for scene n431 than it is for scene xtv8_001.

As for the contrast and noise scores, these are also fairly close to each other in most cases, except for algorithms 4 and 5. For these two algorithms, we also see a marked scene dependence in the contrast scores. When either of these two algorithms is applied to the scenes, scene n431 is judged to have a relatively high contrast, whereas xtv8_001 has a fairly low contrast. For algorithm 3, the opposite is found: contrast is judged much higher (compared to the other algorithms) for xtv8_001 than for n431. This holds for sharpness as well. Algorithm 2, finally, produces the lowest sharpness and contrast in all cases.

We will discuss these findings further in Section 4.

3.2 Multidimensional scaling analysis

In this section, we introduce the multidimensional scaling approach; also see [EMR95] and [KM95b] for detailed examples of this technique. Consider the data for a fixed scene, s. As explained in Section 3.1, four sets of scores on interval scales are available for this scene: $Q_s(i)$, $N_s(i)$, $S_s(i)$ and $C_s(i)$. We now write these scores in a matrix M having 5 rows and 4 columns, in which each row corresponds to an algorithm (i.e., a stimulus) and each column represents a perceptual attribute. The first column contains the quality scores, the second column contains the lack-of-noise scores, the third column contains the sharpness scores, and the fourth column contains the contrast scores, as illustrated below.

$$M = \begin{pmatrix} Q_s(1) & N_s(1) & S_s(1) & C_s(1) \\ Q_s(2) & N_s(2) & S_s(2) & C_s(2) \\ \vdots & \vdots & \vdots & \vdots \\ Q_s(5) & N_s(5) & S_s(5) & C_s(5) \end{pmatrix}$$

To consider a more general situation, we have got a matrix of observations M of dimensions $s \times a$: s stimuli by a attributes. Multidimensional scaling analysis is used to describe both the stimuli and the attributes in a d-dimensional space for some value of d. In terms of matrices, this problem is translated as follows. We want to decompose the $s \times a$ observation matrix M into a stimulus matrix S (dimensions $s \times d$) and an attribute matrix A (dimensions $a \times d$) such that $M = SA^T$. The rows of S can be seen as stimulus positions (vectors) in a d-dimensional space and the rows of A can be seen as attribute vectors in a d-dimensional space. The inner product of row i of S with row j of A would then give the observed score for stimulus i and the attribute j, M_{ij}. Graphically, the scores can be seen as the orthogonal projections of the stimulus vector s_i on the attribute vector a_j, scaled by the length of a_j ; see Figure 5.

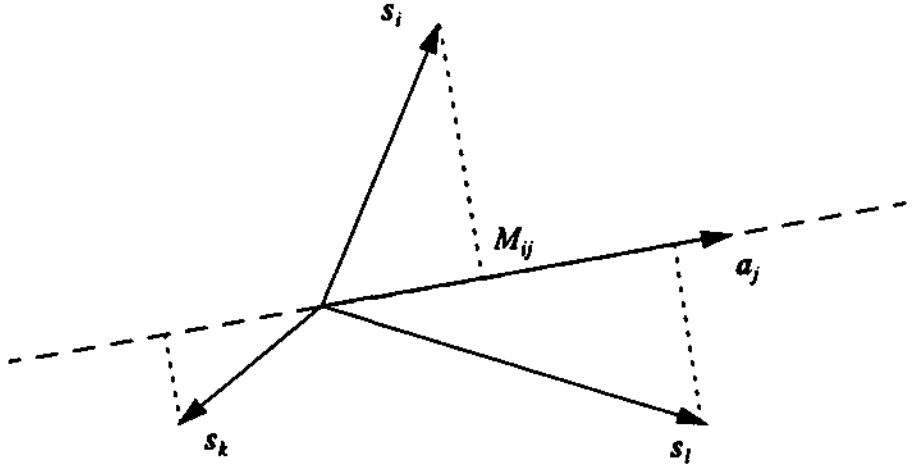


FIGURE 5. Stimuli and attributes as vectors in a 2-dimensional space. The score for stimulus i on attribute j (M_{ij}) can be found by projection of stimulus vector s_i on attribute vector a_j, after normalization of the length of a_j.

Scores projected on a given attribute vector should be interpreted on an interval scale but not on an absolute scale, which means that the scale can be shifted and stretched. This is a consequence of the fact that each attribute was originally ranked on a separate scale. Thus scores '5' for contrast and '7' for sharpness of the same stimulus cannot be compared, but contrast scores '5' and '7' for two different stimuli can. In terms of the graphical representation, all scores that can be compared are projections on a single attribute vector so that they are scaled by the same vector length. Thus lengths of attribute vectors do not matter; only their directions relative to the stimuli matter.

The above mentioned decomposition of matrix M into matrices S and A is only possible if the matrix M has rank at most equal to d (since S and A also have rank at most d). In general, however, the matrix M consisting of real observations will have full rank (i.e., $\text{rank}(M) = \min(s,a)$). Therefore the product of S and A can only be an approximation of M:

$$M = SA^T + E \tag{1}$$

where E is the matrix consisting of error terms. The solution to this problem is explained in Appendix B. The appendix also discusses how to find the "best choice" of the number of dimensions, d. As explained there, d=2 was the best choice for our data. Using two dimensions, 99.7% of the variance could be explained for the case of scene xtv8_001 and 99.5% of the variance could be explained for scene n431.

We solved the problem using the principal components routine in the statistical package SAS. The solution is depicted in Figure 6. In these two plots (one per scene), the length of each attribute vector is proportional to the square of the correlation coefficient between the observed data and the model predictions for that attribute. This is a measure for the goodness of fit. Since correlations cannot be larger than 1, the unit circle plotted in the figure shows the maximum possible vector length. The units on the axes are arbitrary.

It should be noted that the solution found in this way is not unique. In fact, if equation (1) describes one solution, then any invertible $d \times d$ matrix X gives rise to a different solution:

$$M = (SX)(X^{-1}A^T) + E. \tag{2}$$

Hence all stimulus positions and attribute directions are determined up to an invertible linear transformation. Thus no meaning can be attached to angles between vectors: if two attribute vectors happen to be orthogonal (such as contrast and quality for scene n431), this does not necessarily imply that these attributes evoke independent responses.

It is possible to resolve some of this uncertainty in the stimulus configuration by doing additional experiments on dissimilarity judgements. We did such a dissimilarity rating experiment with one subject and four repetitions per pair of stimuli. The task was: to rate the dissimilarity of the pair of stimuli ("how different do they look?") on a scale from 1 to 10. These responses can again be transformed to a stimulus configuration in a multidimensional perceptual space (cf. [GCS89]). In this case, however, the configuration is determined up to an arbitrary *orthogonal* transformation. In such a configuration, angles do have meaning. For our data, the program MULTISCALE (cf. [Ram77]) was used to find 2-dimensional configurations of stimuli for each of the two scenes.

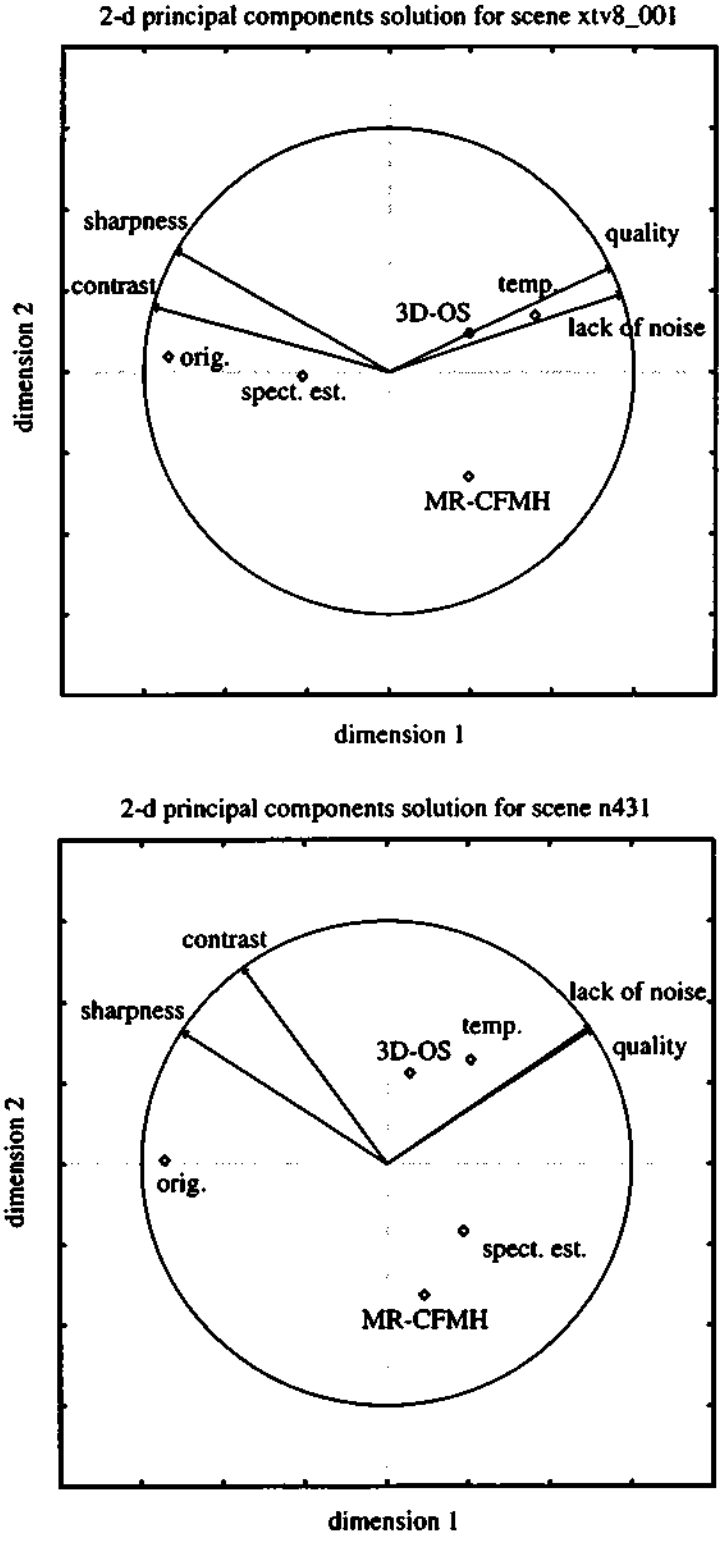


FIGURE 6. Solution of the principal components analysis of the preference data. See text for details.

In this analysis we allowed for a power function to transform the scaled dissimilarities into distances (reminiscent of Weber's law). For the goodness-of-fit measure, we assumed a log-normal error distribution; this means that larger errors in judgements occur when distances are larger. This assumption agrees with general findings in psychophysics.

Let us call the stimulus configuration matrix arising from the principal components analysis (i.e, the preference judgements) S_P. This corresponds to matrix S in equation 1 and 2. Let S_M be the configuration according to the MULTISCALE solution, corresponding to the dissimilarity judgements. Thus S_P is determined up to multiplication with an arbitrary invertible matrix, say X_P, and S_M is determined up to multiplication with an orthogonal matrix X_M (thus $(X_M)^{-1} = (X_M)^T$). If the two stimulus configurations arising from preference and dissimilarity judgements really correspond to one and the same perceptual space, it should then be possible to find X_P and X_M such that

$$S_P X_P = S_M X_M . \tag{3}$$

In practice, such matrices X_P and X_M can be found using a certain goodness-of-fit measure. This approach is implemented in Ramsay's MATFIT procedure (cf. [Ram91]). If we transform the matrix S_P to $S_P X_P (X_M)^T$, the resulting configuration can be compared to the MULTISCALE solution S_M. The comparison is shown in Figure 7, for each of the two scenes. When taking a closer look at the mapping represented by the matrix $X_P(X_M)^T$, it turns out that it can be described as a scaling of the second (y-) axis relative to the first axis, followed by a rotation of the second axis relative to the first axis, followed by a common scaling of both axes and a common rotation. In our case, we found that for xtv8_001, the second axis was scaled by a factor 1.2 and rotated over -69 degrees; the common scaling factor was 4.8 and the rotation was over -172 degrees. For scene n431, the second axis was scaled by a factor of 0.4 and rotated over -39 degrees; both axes were scaled by 7.0 and rotated over -176 degrees. The most noticeable parameter here is the scaling factor 0.4 for scene n431. It implies that a large part of the information (the variance) is shifted to the first dimension, and the second dimension is decreased in importance. The large contribution of the first dimension comes from a large perceived difference between the original scene on the one hand and all processed images on the other hand.

The original plot of the principal components analysis (Figure 6) can now also be transformed: the stimulus positions, given in matrix S_P, are mapped to $S_P X_P (X_M)^T$, and the attribute vectors given in matrix A are changed into $A(X_P)^{-T}(X_M)^T$. This transformation of A gives the new directions of the attribute vectors, but their lengths should be rescaled such that they are still proportional to the squared correlation coefficients, as in Figure 6. This transformed plot, which is more meaningful than the original - the space now being approximately Euclidean - is shown in Figure 8. Again we see a significant difference for scene n431, if we compare this figure to Figure 6: the relative weight of dimensions 1 and 2 has changed. The differences between the two figures are much smaller for scene xtv8_001.

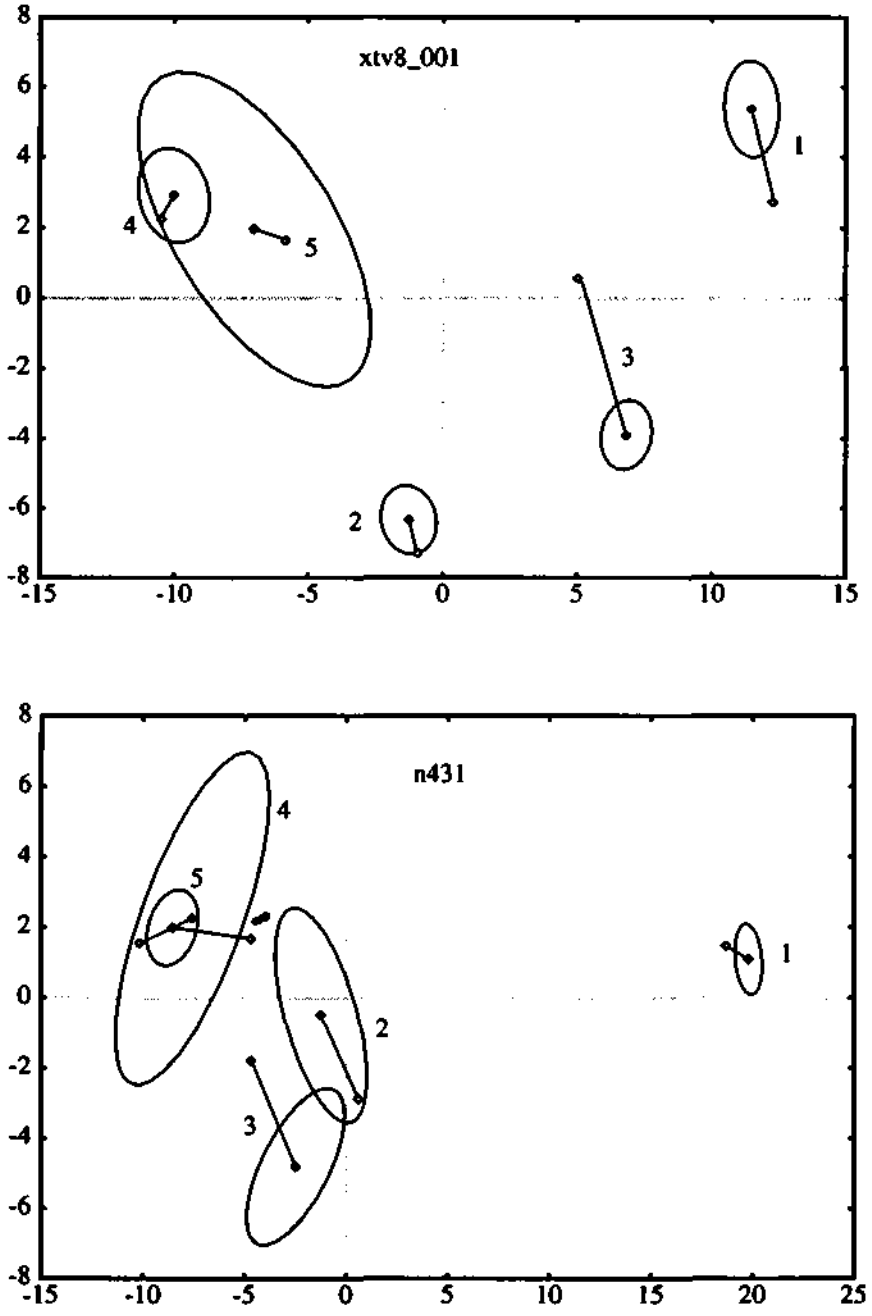
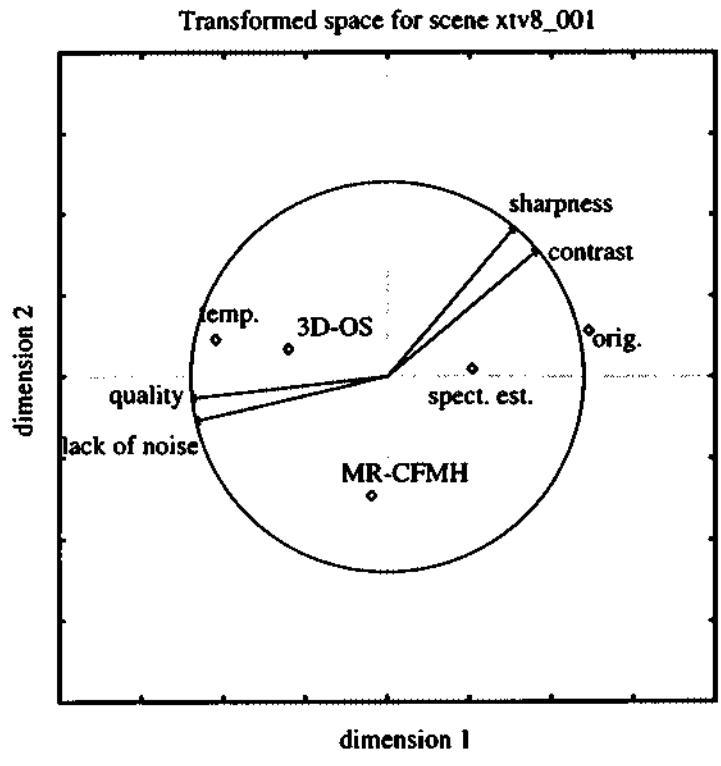


FIGURE 7. Correspondence of MULTISCALE and principal components solutions. The filled symbols represent the stimulus positions in the MULTISCALE configuration and the open symbols are the locations according to the principal components solution. Corresponding stimulus points are joined by straight lines. The ellipses around the filled symbols indicate the region in which the stimulus is located with 95% confidence. Numbers indicate the algorithms (see Table 1).



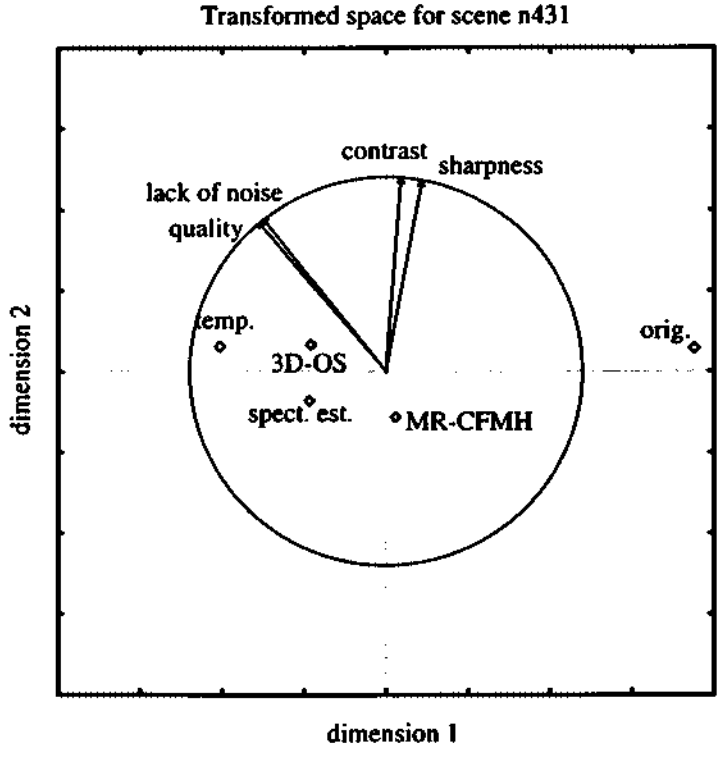


FIGURE 8. Principal components solution after transformation to a "Euclidean" space.

If we interpret the perceptual spaces as shown in Figure 8, we see that the vectors for quality and lack of noise are almost identical. This holds for both scenes. Contrast and sharpness are also very close to each other, for both scenes. The angle between quality and lack of noise on the one hand and sharpness and contrast on the other hand, however, strongly depends on the scene. Whereas high sharpness and contrast seem to be more or less opposed to quality for the xtv8_001 scene (an angle of approximately 140 degrees), the vectors of quality, contrast and sharpness are closer to each other in the case of n431 (an angle of 50 degrees). This can be interpreted as the fact that the noise reduction algorithms generally have a much more detrimental effect on contrast and sharpness for the xtv8_001 scene than for the n431 scene. This may be due to the fact that xtv8_001 had more noise to start with, so that it is more difficult to reduce noise while keeping the sharpness and contrast in this case. For the n431 scene, it is easier to reduce the noise so that in this case, a noise-reduced image even seems to be "clearer" or more "transparent" (terms used by some of the subjects), indicating higher contrast and sharpness.

As for the locations of the stimuli in the perceptual space, we notice that the original scene is always located some distance away from all processed sequences, but this distance is much larger in the n431 case than it is in the xtv8_001 case. Especially the spectral estimation algorithm is still relatively close to the original xtv8_001 sequence, which is due to the fact that the spectral estimation algorithm does not remove as much noise as the other algorithms in this noisy scene. The spectral estimation does a better job for the n431 scene, so this algorithm and the original version of the n431 scene are further apart.

We also notice that the temporal algorithm and the 3-d order statistics algorithm are quite close to each other, for both scenes. In both cases, the temporal, the 3-d order statistics and the original are almost located on a straight line parallel to the x-axis. This could suggest an interpretation for the first dimension: it expresses the effect of the temporal algorithm compared to the unprocessed sequence (i.e., the significant noise reduction achieved for static images). Thus we can say that the 3-d order statistics algorithm is close to the temporal algorithm, but it is shifted a little in the direction of the original.

The second dimension (the y-axis) plays the largest role for the multiresolution-CFMH algorithm. For both scenes, this algorithm is the furthest "down" on the y-axis. It is followed by the spectral estimation algorithm, and the other three algorithms are all located at the high end of the y-axis. The second dimension thus mainly describes the effect of the multiresolution algorithm as compared to either the original or the temporal algorithm; an effect which is related more to contrast and sharpness than to noisiness.

4 Discussion

4.1 Subjects' remarks

In this section, we describe the remarks made by the subjects when they were presented with the stimuli. We distinguish remarks about the scenes themselves, about the perceptual attributes that had to be judged, and about the algorithms. Some of these remarks help to explain the findings in the previous section, but others contradict these. They also suggest some directions for improvement, mentioned in Section 4.2.

Scenes

One subject remarked that sequence xtv8_001 was "ugly" to start with, because of the large amount of noise and low contrast. There are no indications that the extra noise from the video recording (see Section 2.3) looked strange to the subjects. Several subjects remarked that images usually did not contain global motion in this type of examination. This lead them to consider mainly the static part of the sequence when judging the quality, sharpness, noisiness and contrast.

For sequence n431, the remarks concentrated on the fact that this sequence was very short and the motion contained in it was too fast and too short for them to be able to judge it reliably. Thus also in this case, the judgements were predominantly based on the static phase of the sequence.

The fact that the static parts of the two sequences played a larger role than the dynamic part in the subjects' judgement may have had a large effect on the outcome of the experiments. The temporal and 3-d order statistics algorithms were rated so high because they produced a beautiful result when the images were static. When the images started to move, these two algorithms caused a very annoying noise breakthrough, especially for the xtv8_001 scene. Although this was mentioned by almost all of the subjects, most of them still preferred the temporal and 3-d order statistics algorithms in the split-screen experiment because they put more weight on the static parts.

Perceptual attributes

We now go into the remarks about the different perceptual attributes judged by the subjects - sharpness, contrast, lack of noise and quality - and the image details they used to judge these attributes by. The images in Appendix D can be used as an illustration of these details.

Sharpness was sometimes difficult to judge, according to the subjects. Again this is due to the difference between static and moving images: a sequence could look very sharp when static and be blurred when moving. Also, the various algorithms did not cause very different appearances in sharpness. It was remarked that the amount of noise in an image could influence the impression of sharpness: the presence of high frequency noise could help the subject to focus better and thus elevate the perceived sharpness of the image.

Details used to assess the sharpness were: the guide wire (used by all subjects), the catheter and the rims of the vertebrae in the xtv8_001 sequence; in isolated cases also the thin vertical lines between vertebrae were used. For the n431 sequence, the sharpness was mainly judged by the folds in the colon wall. Many subjects had concentrated on two thin folds near the top part of the colon, because these folds could in some cases disappear when the colon started to move. The subjects also looked at the contour of the colon and the small diverticuli to decide on the sharpness.

Contrast was even more difficult to assess than sharpness. One subject did not see any difference whatsoever between the contrast of the differently processed sequences, and another one was only able to judge it in the xtv8_001 scene. The problem with the n431 scene was that it contained only a few different grey values: black where the barium was, mid-grey in the background and in the part of the colon where there was no solid barium, and light grey in a small part of the colon apparently containing air. In this scene, contrast could either be judged as the blackness of the barium, or as the dynamic range (the difference between the black and the light grey), or as the amount of variance in the mid-grey area: the visibility of details in the colon. The xtv8_001 scene contained many more grey shades and its contrast was easier to judge for most subjects. Their judgement was based on the darkness of the catheter, the cloud of contrast liquid, and the rims of the vertebrae. A few subjects also took the low contrast details in the background into consideration.

For both scenes, contrast could be affected by noise in different ways. When low contrast details were used to judge the contrast by, a high amount of noise corresponded with a low contrast judgement, because the low-contrast details were drowned in the noise. When the "global" contrast was judged (related to the dynamic range of the whole image), on the other hand, the presence of noise could enhance the contrast impression. This can be explained by the bright and dark noise peaks which occasionally occurred in the sequence and which added to the perceived dynamic range. This is similar to the increased sharpness which seems to accompany fine-grained noise.

Noisiness was usually easy to judge for the subjects. This also follows from the perfect agreement between subjects (see Figure 1 and Figure 2) when it comes to assessment of the noise. The only difficulty that some subjects had was with the different appearance of the noise. When they had to compare one of the spatial algorithms with one of the temporal ones, the two types of noise looked very different: the spatial algorithm produced a relatively high noise level which was constant over the whole sequence, while the temporal algorithm produced only a little noise (looking like fixed pattern noise) during the static parts but a very high amount of noise during motion. In this case the subjects said to judge the *annoyance* of the noise rather than the *amount* of noise.

The quality of the images could depend on various things. Of course the noise, sharpness and contrast played a role, but many other aspects of images were mentioned. To wit: noise breakthrough at the onset of motion, causing things like holes in the guide wire; the appearance of noise near sharp edges ("edge business"); the visibility of patterns in the noise, as if the image was seen though a dirty window; artefacts like "blotches" or "patches", or just the fact that an image looked "artificial", as if it were painted with watercolours. Which remarks corresponded to which algorithm is discussed below.

Image details that were used to judge the quality of the sequences were usually related to the diagnostic or interventional task for which these sequences would be used. For the xtv8_001, sequence, the criteria were: the visibility of the guide wire and catheter (specifically the tip of the catheter), and how well the cloud of contrast liquid could be seen spreading out. For n431, the quality depended on the visibility of structure in the colon, on diagnostically important details like the contours and folds, and also on the "3-d impression". The latter remark was explained as the ability to see how the bowels were actually located inside the patient: how the curves of the bowels were projected on top of each other.

Algorithms

Finally we discuss the remarks related to the different algorithms. Obviously, the original unprocessed sequences were recognized as the most noisy ones. In general, they were also considered to have the lowest quality. The opinion on algorithm number 2, the MR-CFMH algorithm, was that it achieved some noise reduction, but still a fair amount of noise was left in the image (particularly for the xtv8_001 scene). Sequence xtv8_001 also looked a little "flat" or "hazy" when processed with this algorithm. With respect to scene n431, there was some mention of motion blur (which also occurred in the original sequence) and noise breakthrough. It is interesting to note that there were only few remarks about the unsharpness of the sequences processed with this algorithm, although the 2AFC experiment clearly showed that this algorithm produced the least sharp images of all algorithms, for both scenes (see Figure 4). This implies that the differences in sharpness are in fact small, and are only noticeable when two versions are directly compared.

Algorithm 3 (spectral estimation) was considered by some people to be worse than algorithm 2 for the xtv8_001 sequence, in the sense that this algorithm left somewhat more noise. Yet algorithm 3 produced a sharper image, and the noise was more fine-grained, which a few of the other subjects preferred. For scene n431, algorithm 3 was generally preferred over algorithm 2. Algorithm 3 reduced more of the noise but it did not introduce more unsharpness or contrast loss than algorithm 2 did. Artefacts were never mentioned for this algorithm.

Algorithm 4 - the temporal anisotropic diffusion - was thought to be the best algorithm when the image was standing still: it achieved a very good noise reduction while preserving the sharpness and contrast. One of the subjects complained that the xtv8_001 scene processed with this algorithm looked unnatural because of the heavy noise reduction: patchy, as if painted with watercolours. When the image started to move, most of the subjects (seven out of ten) noted the severe noise breakthrough near the edges that was introduced by this algorithm. This was more apparent in the xtv8_001 sequence than in the n431, probably because more noise was breaking through in the xtv8_001 scene. Some of the subjects even feared that this will make the algorithm unacceptable to certain radiologists. Note the controversy between these remarks and the findings in the 2AFC experiments, in which the temporal algorithm turned out as the best one. The remarks about a "dirty window" and a noise pattern were also made for this algorithm. They applied mainly to the n431 scene. This effect may be explained from the so-called "structured noise": certain small, seemingly random variations in the image which do not vary in time.

This type of "noise" is not removed by the temporal algorithm, because it is present in every frame and thus it is considered as signal. Only the time-varying noise is removed from the images, causing the stationary structured noise to become more conspicuous.

Lastly, the fifth algorithm (3-d order statistics) gave rise to remarks that were very similar to those for the temporal algorithm. When comparing the two algorithms, it was generally acknowledged that the noise breakthrough and the dirty window effect were less bad in the 3-d order statistics algorithm, but at the same time the 3-d order statistics algorithm left a little more noise in the images. The higher noise level may explain the observation that the breakthrough and the structured noise artefacts were less visible, because the noise is masking some of these artefacts. Most remarks indicated that the less artificial look of algorithm 5 was preferred over algorithm 4, but again this was contradicted by the results of the 2AFC experiments.

4.2 Conclusions and recommendations

In this final section we summarize our conclusions, based on both the experiment and the subjects' remarks. We also present some recommendations for improvement of the algorithms and for future evaluations.

First of all, all noise reduction algorithms perform better than the original ("algorithm 1"). The perceived amount of noise is highest in the originals, and the quality is lowest. On the other hand, the contrast and sharpness are also the highest for the original, indicating that noise reduction always introduces some degree of contrast loss and unsharpness. It is at this point not clear whether this is an undesirable artefact of the algorithms, or whether the presence of noise in itself enhances the perceived sharpness and contrast. There is some evidence of this, as was shown in some IPO studies ([RV93], [KM95a]): for slightly blurred images, the perceived sharpness can be higher when some white Gaussian noise is added to the images. The question whether this effect can explain our data should be studied separately, e.g. by comparing a "noise-free" image (obtained by temporal averaging of the frames in a static phantom sequence) to the original noisy image and comparing this to a version of the image treated with one of the noise reduction algorithms.

From the experiments, we may conclude that the quality judgements are almost identical with the lack-of-noise ratings. Similarly, we found that contrast and sharpness ratings are very close to each other. However, we should keep in mind that this holds after pooling the data over all subjects. When looking at the individual results, we see that the quality and lack-of-noise ratings are still very close, but the sharpness and contrast scores show more spread.

The observation that quality is analogous to "lack of noise" is only partly supported by the remarks made by the subjects. They agreed that the amount of noise was important for the quality, but they seemed to take only the static parts of the images into account during the experiment, whereas they considered the moving parts as well when they were making their remarks.

Marked differences exist between the results of different subjects. It was to be expected that differences in taste would exist, which would result in a spread in the "quality" results, but we also found differences in the "sharpness" and "contrast" results. There are two possible explanations for this. The first reason could be that different people used different aspects of the images to judge the sharpness and contrast on. This explains part of the results for contrast (e.g., it makes a difference whether the blackness of the barium or the low-contrast details in a bright background are judged), but it is much less likely that this also holds the sharpness results. The second explanation is more plausible: namely that the different algorithms are very similar with respect to both contrast and sharpness, so that it was sometimes difficult for the subjects to point out the image with the higher sharpness or contrast. This introduces "noise" in the results, since every pair of images was compared only four times per subject. This can be interpreted as a good sign, because the algorithms are not supposed to alter the sharpness or the contrast of the sequence to which they are applied. Note, however, that the spread in the contrast results is larger than the spread in the sharpness results. This suggests that the first explanation (using different features for the judgement) may play an additional role for the contrast results but not for the sharpness results.

The inter-subject spread in the quality results is smaller than the spread in sharpness or contrast, at least for most of the subjects. Following the above reasoning, this could imply that differences in quality were more pronounced than differences in sharpness or contrast, so that the quality data are less corrupted by "noise". This is confirmed by the remarks of the subjects. Still there are some obvious taste differences, which cannot be neglected. The "quality" correlation tree in Figure 2 clearly shows this. We therefore recommend that any algorithm to be implemented in a real system should provide a "customization" feature: the user should be allowed to vary certain parameters in the processing to tune the appearance of the images to his or her liking; e.g., the amount of noise to be reduced, or the algorithm's sensitivity to motion. It should be possible to vary these parameters in an easy and intuitive way. This aspect of customization is a topic for further study.

Related to customization is the tuning of the algorithms to different fields of application or different fluoroscopic techniques (e.g. pulsed fluoroscopy at different frame rates, trace subtract, frame grab...). This has not been investigated in this study, but there are indications that different applications might benefit from different settings of the noise reduction algorithms. Such settings could be provided as default settings, for instance in the APR or in "fluo flavours".

Our conclusions with respect to each of the algorithms are as follows. The temporal algorithm emerged as the best one in the 2AFC experiments. However, most of the subjects strongly objected to the global noise breakthrough when they were asked to express their opinion about each of the algorithms. Some of them even considered this as inadmissible. Since this algorithm gave excellent results on static images, it seems worthwhile to try to improve its performance during motion, by incorporating motion estimation and motion compensation. It was mentioned by one of the subjects that noise breakthrough is the most objectionable when it occurs as a consequence of global motion (i.e., noise breaking through simultaneously at different edge locations in the image), and noise breakthroughs due to local motion may be less bad. This indicates that a relatively simple global motion

estimation routine could be sufficient to eliminate the worst noise breakthrough effects. The alternative way to get rid of the noise breakthrough is to reduce less of the noise; thus, to leave a higher level of noise throughout the image sequence so that the artefact is masked. Of course this deteriorates the performance of the algorithm during static phases of the image sequence.

We also found that the 3-d order statistics algorithm was quite close in performance to the temporal algorithm, both with respect to the amount of noise reduction as to the contrast and sharpness appearance. Even the type of artefacts was very similar in the two algorithms, although these were less noticeable in the 3-d order statistics algorithm. As said before, the 3-d order statistics algorithm reduced a little less noise and maintained a somewhat higher sharpness and contrast. The experiments implied that the 3-d order statistics algorithm was slightly worse than the temporal one, but the subjects' remarks sometimes indicated the opposite. The fact that the very simple 3-d order statistics algorithm is perceptually so close to the more complex recursive temporal filter makes it an interesting candidate to study further, perhaps improving its sensitivity to motion.

The multiresolution-CFMH algorithm left more noise in the images than the temporal and 3-d order statistics algorithms. It also produced the lowest sharpness and contrast of all algorithms. We conclude this mainly from the 2AFC experiments. The observation that the images were still fairly noisy was strengthened by subjects' remarks; the pronounced results for sharpness and contrast, however, were not supported by the (few) remarks about these attributes. It is of course conceivable that the disappointing results of this algorithm are due to a suboptimal choice of the algorithm parameters (such as the number of iterations or the amount of contrast enhancement).

The spectral estimation algorithm, finally, behaved differently for the two scenes. For the xtv8_001 scene, it achieved relatively little noise reduction - even less than the multiresolution CFMH algorithm - but the contrast and sharpness were higher than for any of the other algorithms ("algorithm 1" excluded: the original still had the highest contrast and sharpness). Better noise reduction was obtained in the n431 sequence, but here the contrast and sharpness were low. The subjects' remarks agreed with the experimental results. The absence of artefacts is certainly an advantage of this algorithm, which makes it interesting to investigate whether a different tuning of the parameters could improve the performance for the noisiest scene (xtv8_001).

As we have seen, all algorithms introduce some amount of contrast loss and blur. If the noise reduction is "good enough", one could compensate for the contrast and sharpness loss by incorporating some contrast and/or sharpness enhancement in the noise reduction algorithm. In fact, this was already done to some extent in the multiresolution CFMH algorithm. The amount of enhancement would have to depend on the noise level and possibly also on factors like the application area and personal preference.

We end with a few remarks about practical aspects of the evaluations themselves, which should be considered when future evaluations have to be carried out. One issue is the choice of scenes. If effects of motion (like noise tails) play a role, then it is important to use sufficiently long sequences. The n431 sequence, consisting of 47 frames, was not long

enough for this purpose. In the case of sudden jerky motion as occurring in scene n431, it could also be considered to use Divise's "shuttle" option instead of "repeat". Although the shuttle option shows temporal effects in the wrong order, half of the time, it gives a much more restful appearance of the moving sequence and subjects would be less distracted by the jump every time the sequence starts over again.

A second remark deals with the questions the subjects had to answer. In the 2AFC experiment, we concentrated on four perceptual attributes. Although the attributes contrast, sharpness and lack of noise seemed to explain most of the quality ratings, it turned out that very different image aspects like the annoyance of noise breakthroughs and "dirty window" effects were important as well. This could only be gathered from the informal remarks made by the subjects after the 2AFC experiment was finished. Therefore it seems advisable to include questions about the visibility of such artefacts in a more formal way; e.g., to ask "which image suffers more from motion artefacts, left or right?".

An final issue is the monitor used for the evaluations. It is hard to say how much our results are influenced by the physical characteristics of this particular monitor. During the development of the algorithms, different monitors were used which had a better modulation transfer function (MTF). Certain aspects of images, like high frequency noise in a 1024 x 1024 pixel image, would look different on such a monitor. We tried to prevent such dependencies on the MTF by using images of 512 x 512 pixels, such that the images did not contain very high frequency information. The grey-value-to-luminance curve of the monitor will also affect the results; not only for contrast, but also for the visibility of noise. The same can be said about the amount of ambient light and the viewing distance. Our main reason for choosing this monitor with the given setting and viewing conditions was that this was the same type of monitor, and viewed in comparable conditions, as can be found in clinical situations: the environment in which the final product will be used.

Acknowledgement

I would like to thank the members of the "Image processing for fluoroscopy" project team for providing me with the image material, and for helpful discussions on the preliminary results of this work. Thanks are also due to the people at the PMSN X-ray Predevelopment group for enabling me to run the experiments on their ISP system. Finally, I am deeply indebted to the nine application experts who participated in the experiment, for their patience and endurance.

References

[AK95a] T. Aach and D. Kunz, "Spectral techniques for quantum noise reduction in X-ray fluoroscopy images", Laborbericht Nr. 1128/95, PFL-A, 1995.

[AK95b] T. Aach and D. Kunz, "Efficient multiresolution/multiscale nonlinear filters for X-ray quantum noise reduction and X-ray image enhancement", Laborbericht Nr. 1144/95, PFL-A, 1995.

[AN93] A.J. Ahumada and C.H. Null, "Image quality: a multidimensional problem". In: A.B. Watson (Ed.), *Digital images and human vision.* MIT Press: London, 1993

[ACSK90] R.L. Arenson, D.P. Chakraborty, S.B. Seshadri and H.L. Kundel, "The digital imaging workstation", Radiology, vol. 176, pp. 303-315, 1990.

[DR83] P. Dunn-Rankin, *Scaling methods.* Lawrence Erlbaum Acc.: Hillsdale, NJ, 1983.

[EY36] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank", Psychometrika, vol. 1, pp. 148-158, 1936.

[EMR95] B. Escalante Ramírez, J.B. Martens and H. de Ridder, "Multidimensional characterization of the perceptual quality of noise-reduced Computed Tomography images", Journal of Visual Communication and Image Representation, vol. 6, no. 4, pp. 317-334, 1995.

[Flo95] R. Florent, "Temporal recursive filter for noise reduction". LEP report, to appear.

[GCS89] P.E. Green, F.J. Carmone, Jr and S.M. Smith, *Multidimensional scaling - concepts and applications.* Allyn and Bacon: Boston, 1989.

[GL83] G.H. Golub and C.F. van Loan, *Matrix computations.* p. 19-20. North Oxford Academic: Oxford, 1983.

[KM95a] V. Kayargadde and J.B. Martens, "Perceptual characterization of images degraded by blur and noise: model". Submitted to J. Opt. Soc. Am., 1995.

[KM95b] V. Kayargadde and J.B. Martens, "Perceptual characterization of images degraded by blur and noise: experiments". Submitted to J. Opt. Soc. Am., 1995.

[Kun86] H.L. Kundel, "Visual perception and image display terminals". Radiologic Clinics of North America, vol. 24, no. 1, pp. 69-78, 1986.

[MN75] H. Marmolin and S. Nyberg, "Multidimensional scaling of image quality". FOA report C-30039-H9, Swedish National Defense Research Institute, Stockholm, 1975.

[Ove93a] W.M.C.J. van Overveld, "The effect of gamma on subjective quality and contrast for X-ray images". IPO report no. 907, 1993.

[Ove93b] W.M.C.J. van Overveld, "A study on X-ray image quality based on interviews with radiologists in the USA". IPO report no. 921, 1993.

[Ove95a] W.M.C.J. van Overveld, "Contrast, noise, and blur affect performance and appreciation of digital radiographs". IPO Manuscript no. 1091; to appear in Journal of Digital Imaging, November 1995.

[Ove95b] W.M.C.J. van Overveld, "Image quality issues for fluoroscopy assessed through interviews". IPO report no. 1056, 1995.

[Ram77] J.O. Ramsay, "Maximum likelihood estimation in multidimensional scaling". Psychometrika, vol. 42, pp. 241-266, 1977.

[Ram91] J.O. Ramsay, "MULTISCALE manual", McGill University, 1991.

[RV93] M. de Ridder and J.P.A. Verbunt, "De invloed van ruis en onscherpte op beeldkwaliteit". IPO report no. 956, 1993 (in Dutch).

[Rou92] J.A.J. Roufs, "Perceptual image quality: concept and measurement". Philips Journal of Research, vol. 47, pp. 35-62, 1992.

[SL93] P.J. Stemkens and H.A.P. van de Loo, "Digital Video Simulation Facility - Unix Divise user manual". Philips Research Laboratories, Eindhoven, Report number RWR-028-PS-93024-ps, 1993.

[Tor58] W.S. Torgerson, *Theory and methods of scaling*. John Wiley and Sons: New York, 1958.

[Tre68] H. L. van Trees, *Detection, Estimation, and Modulation Theory. Part I: Detection, Estimation, and Linear Modulation Theory.* John Wiley and Sons: New York, 1968.

Appendix A: Three methods of transforming ranking data to an interval scale

Variance stable rank method

According to this method, the scale value of a stimulus equals the number of times the stimulus is preferred to any other stimulus, divided by the total number of times the stimulus occurred in a comparison. This method gives scale values between 0 and 1, but of course this range can be changed by any linear transformation. Thus, in our application, the scale value for "general preference" of algorithm j, $Q_s(j)$ is derived as

$$Q_s(j) = a \cdot \frac{\sum_i Q_s(i,j)}{\sum_i Q_s(i,j) + \sum_i Q_s(j,i)} + b \qquad (4)$$

where the constants a and b are chosen such that the scores range from 1 to 10. Similar equations are found for the noise, sharpness and contrast scales $N_s(j)$, $S_s(j)$ and $C_s(j)$.

Thurstone case V method

According to Thurstone's model (cf. [Tor58]), stimuli evoke sensations on a psychological scale which is assumed to be an interval scale. The sensations arising from a single stimulus follow a normal distribution. In case V, it is assumed that the distributions for different stimuli are independent and share the same standard deviation, σ. The mean of the distribution for stimulus i is called μ_i. Thurstone's case V model can be written as

$$\mu_j - \mu_i = z_{ij}, \qquad (5)$$

in which z_{ij} is the normal deviate corresponding to the theoretical probability that stimulus j is judged higher than i. If we apply this to our 'quality' data for a fixed scene s, we can estimate this theoretical probability, say $P_s^*(i,j)$, from the observed proportion of times stimulus j was preferred over i:

$$P_s^*(i,j) = \frac{Q_s(i,j)}{Q_s(i,j) + Q_s(j,i)}. \qquad (6)$$

From this, we derive the estimates for z_{ij} - denoted z_{ij}^* - via the cumulative normal distribution function. After reordering the stimuli in increasing order (i.e, the order found through a computation similar to (4)), the matrix containing the numbers z_{ij}^* can be used to compute distances between consecutive stimuli. The distance between stimulus k and stimulus $k+1$, d_k, is computed as

$$d_k = \frac{1}{n}\left(\sum_i z_{i,k+1}{}^* - z_{ik}{}^*\right) \tag{7}$$

where n is the number of stimuli (or rather, the number of 'nonempty cells', in the case of missing data; see [Tor58] for details). The interval scale can then be computed by fixing an arbitrary starting point for μ_1 and computing all other μ_k through

$$\mu_{k+1} = \mu_k + d_k. \tag{8}$$

Again, a linear transformation can be applied to the μ_k to obtain the range from 1 to 10. Interval scales for lack of noise, sharpness or contrast are derived analogously.

Maximum likelihood method

This method uses the same model as the Thurstone method (equation (5)), but it applies a more sophisticated model to estimate the parameters of the model (i.e., the average stimulus locations) from the observations, namely maximum likelihood estimation. Maximization of the log-likelihood function is done using an iterative procedure called the "scoring" method.

Although this parameter estimation method is more complex that the straightforward Thurstone method, it has the advantage that it also gives an indication for the accuracy of the parameter estimates. The accuracy is derived from the estimated variance-covariance matrix of the model parameters, which is the inverse (or the Moore-Penrose pseudo-inverse) of the Fisher information matrix (containing second derivatives of the log-likelihood function w.r.t pairs of parameters). The theory behind this can be found in e.g. section 2.4 of [Tre68].

Appendix B: Theory of multidimensional scaling analysis

In this appendix, we show the solution to the approximation problem mentioned in Section 3.2, Equation (1):

$$M = SA^T + E \tag{9}$$

For a least-square solution to this problem, $\|E\|$, the Frobenius norm of E (the square root of the sum of all squared matrix elements) is minimum. This problem can be solved by so-called Eckart-Young decomposition ([GL83], [EY36]) in the following way. Let

$$M = U\Sigma V^T \tag{10}$$

be the singular value decomposition of M, in which U and V are orthogonal matrices (U of dimension $s \times s$ and V of dimension $a \times a$) and Σ is an $s \times a$ matrix with $\sigma_1, \sigma_2, \ldots, \sigma_p$ on the diagonal ($p = \min(a,s)$) and zeroes in all other positions. The $\sigma_1, \sigma_2, \ldots, \sigma_p$ are the singular values of M in decreasing order. Eckart and Young have shown that the least-square approximation of M by a matrix of rank d is given by

$$M = U_d \Sigma_d V_d^T + E \tag{11}$$

in which U_d is an $s \times d$ matrix containing the first d columns of U, Σ_d is the $d \times d$ diagonal matrix containing the first d singular values of M ($\Sigma_d = \mathrm{diag}(\sigma_1,\sigma_2,\ldots,\sigma_d)$), and V_d is the $a \times d$ matrix containing the first d columns of V.

Thus the solution of our approximation problem $M = SA^T + E$ is found by taking $S := U_d$ and $A := V_d\Sigma_d$ or alternatively, $S := U_d\Sigma_d$ and $A := V_d$, depending on whether the stimulus or the attribute vectors are scaled to unit length. This solution is used in principal component analysis (see [AN93], the description of MDPREF in [GCS89], or the PRINQUAL routine in the statistical package SAS). The analysis allows for a linear transformation of the input data (assuming this is on an interval scale) prior to finding the principal components. The transformation is chosen such that the error to the model ($\|E\|$) is smallest.

The problem of choosing the "right" value of d is solved by studying the singular values of M. For a d-dimensional solution, it can be shown that the error to the model equals

$$\|E\| = \sqrt{\sigma_{d+1}^2 + \sigma_{d+2}^2 + \ldots + \sigma_p^2} \tag{12}$$

so that the amount of explained variance increases with the number of singular values included in the model. If the singular values satisfy $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_d > \sigma_{d+1} \geq \ldots$ 0 where σ_d is significantly larger than 0 (i.e., the dth dimension still explains a significant proportion of the variance) but σ_{d+1} is close to 0, then it makes sense to approximate the data by a d-dimensional solution. In our case, we found for the data of scene xtv8_001 that the cumulative percentage of explained variance equalled 84.62 for one dimension, 99.74 for two dimensions and 99.98 for three dimensions; so a two-dimensional solution seemed appropriate. For scene n431, these three percentages were 60.32, 99.48 and 100.00, respectively. Again we opted for a two-dimensional solution.

Appendix C: Monitor characteristics

The grey-value-to-luminance curves shown below were measured on two test patterns each containing 16 square patches of linearly increasing grey values. In the test pattern indicated by "bright background" (see plots) these patches are embedded in a background of maximum intensity. In the test pattern labelled by "dark background", the patches occur on a black background.

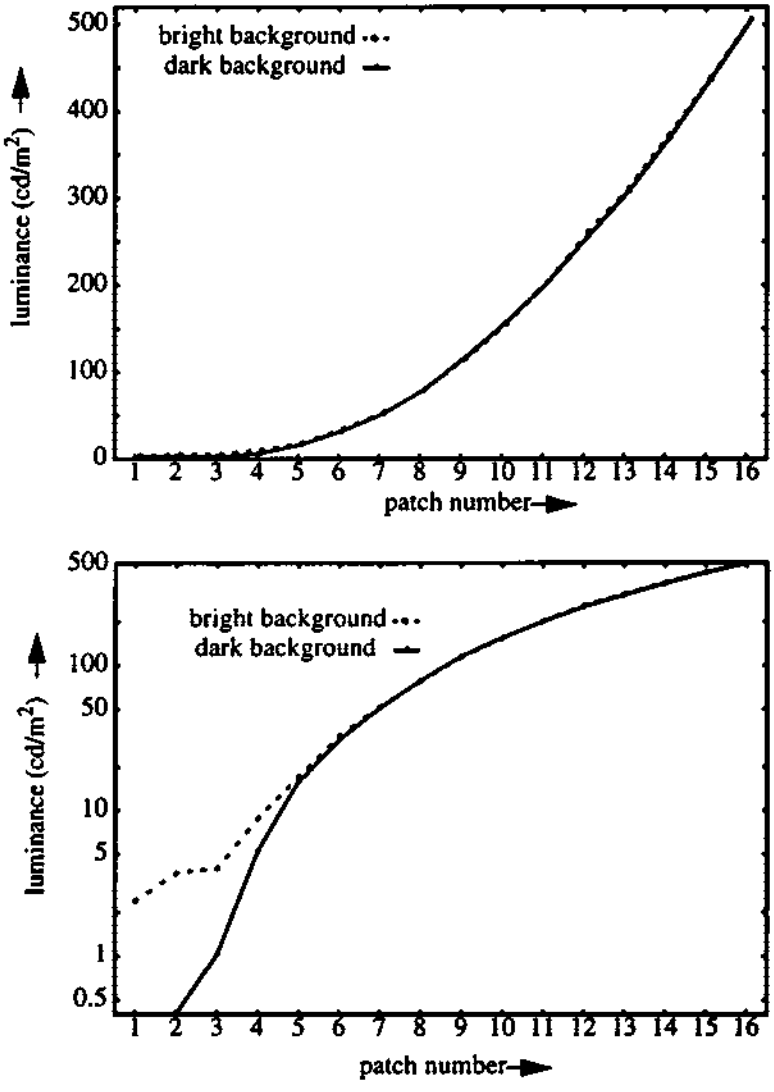


FIGURE 9. grey-value-to-luminance curves with luminance plotted on linear and logarithmic scales.

The luminance of each patch was measured on the screen using a luminance meter of the type Mavo monitor (no. XSB-PST 35048). It can be seen that there is a good correspondence between the measurements for the two test patterns, except for the three or four darkest patches. For those patches, luminances measured on the bright background are about 2 to 3 cd/m^2 higher than those for the dark background. This is probably due to stray light influencing the measurements.

Appendix D: Sample images

The two images below are frames taken from the sequences used in the experiments.

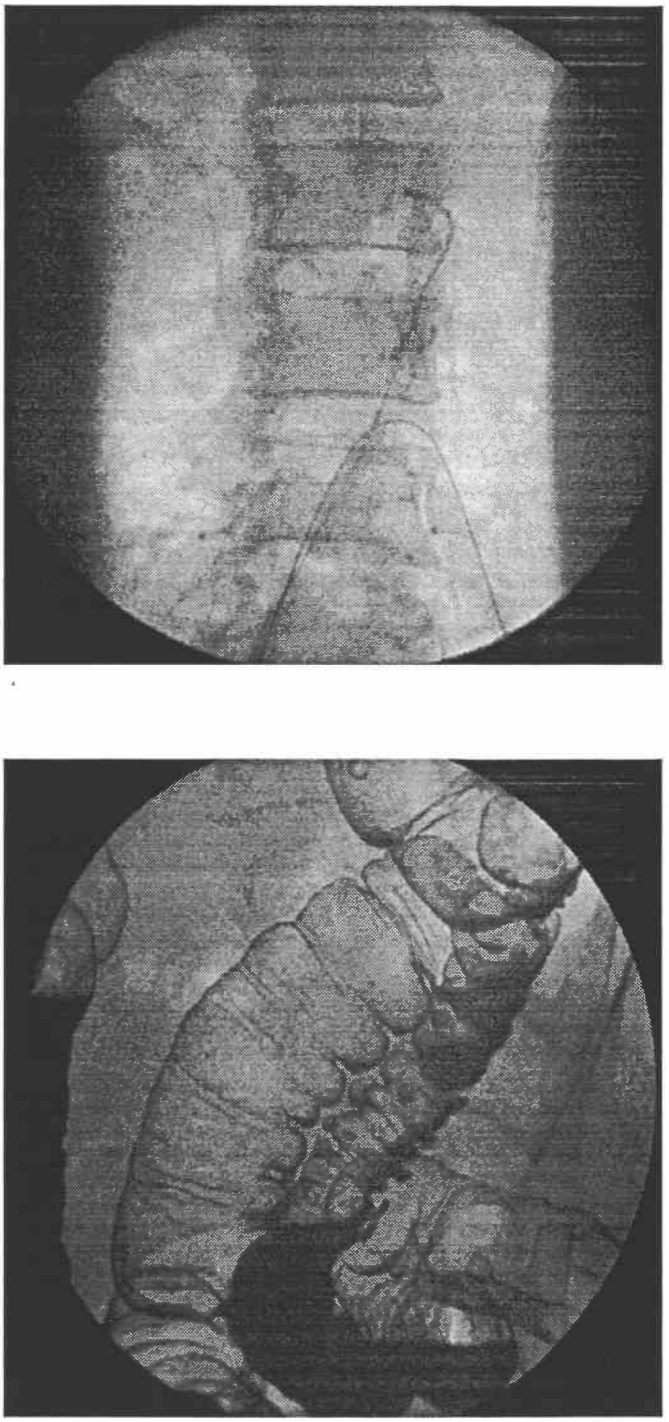


FIGURE 10. Frames from the unprocessed sequences xtv8_001 (top) and n431 (bottom).