14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

PSYCHO-VISUAL QUALITY ASSESSMENT OF STATE-OF-THE-ART DENOISING SCHEMES

Ewout Vansteenkiste[1], *Dietrich Van der Weken*[2], *Wilfried Philips*[1], *Etienne Kerre*[2]

[1]Image Processing and Interpretation Group, TELIN Dept., Ghent University
[2] Fuzziness and Uncertainty Modeling Group, Dept. of Applied Mathematics & Computer Science, Ghent University
Sint-Pietersnieuwstraat 41, Ghent, Belgium
ervsteen@telin.ugent.be
http://telin.ugent.be/IPI/

ABSTRACT

In this paper we compare the quality of 7 state-of-the-art denoising schemes based on human visual perception. 3 of those are wavelet-based filter schemes, 1 is Discrete Cosine Transform-based, 1 is Discrete Fourier Transform-based, 2 are Steerable Pyramid-based and 1 is Fuzzy Logic based. A psycho-visual experiment was set up in which 37 subjects were asked to score and compare denoised images coming from 3 different scenes. A Multi-Dimensional Scaling framework was then used to process the data of this experiment. This lead to a ranking of the filters in perceived overall image quality. In a follow-up experiment other attributes such as the noisiness, bluriness and artefacts present in the denoised images allowed us also to determine why people choose one filter over the other.

1. INTRODUCTION

Denoising has been a hot topic for many years in different image processing and analysis tasks, f.i. in image restoration or as a preprocessing step to segmentation. Multiple advanced schemes have been presented in recent literature using locally adaptive spatial filters in a multi-resolution representation [9, 8, 11], shape-adaptive transforms [4], block-matching with 3D transforms [1], Steerable Filter Pyramid based [10, 5] or Fuzzy Logic [13] techniques.

All of these tend to suppress the noise present while preserving as much image content, structures and detail information as possible. Different well-known measures such as the Root Mean Square Error (RMSE) or Peak Signal-to-Noise Ratio (PSNR) are commonly used to compare how well the different filters perform. Although these are good measures to determine a relative distance, for instance to the original noise-free image (if provided), and accordingly to rank the filters, what do these differences tell us about the overall image quality since they don't incorporate human visual information?

Different alternative measures have been proposed tending to incorporate this kind of knowledge, for example the fuzzy similarity measures described in [14]. Yet for certain purposes, f.i. when image distortions become too small to be well captured by any instrumental measure, a better approach is to determine a ranking solely based on human visual perception, through some psycho-visual experiment [7].

Here, we will perform our own experiment on 7 state-of-the-art denoising schemes, trying to rank the filters in perceived overall image quality and to determine why our subjects prefer one filter over the other.

The Multi-Dimensional Scaling (MDS) framework priorly used by Martens et al. [3] will be used to process the data from the experiment. The rationale underlying this framework is twofold. First, the concept of "homogeneity of perception" should hold, meaning that different subjects are able to reach one common conclusion, f.i. on overall image quality. Secondly, the concept of overall image quality is rarely one-dimensional, meaning that different attributes such as noise, blur or artefacts all can influence the perceived quality.

In the next sections we will first present a brief overview of the different denoising schemes, Section 2, then elaborate on our psycho-visual experiment, Section 3, before turning to the results and conclusions in Sections 4 and 5.

2. DENOISING SCHEMES

From recent literature the following denoising schemes were selected based on good overall performances. For technical details we refer to the papers in the references:

- **The GOA filter** [13]: A two-step filter where first a fuzzy derivative for eight different directions is computed which is then used to perform a fuzzy smoothing by weighting the contributions of neighboring pixel values. Both stages are based on fuzzy rules using membership functions.
- **The SA-DCT filter** [4]: The Shape-Adaptive DCT scheme uses an overcomplete transform-domain filter in conjunction with the anisotropic LPA-ICI technique, which - for every location in the image - adaptively defines an appropriate shape for the transform's support.
- **The 3D-DFT filter** [1]: The block-matching and 3D filtering approach exploits the possible correlation among similar blocks within an image by filtering in the 3D-transform domain. The third dimension corresponds to stacking together the blocks which are matched as similar.
- **The ProbShrink filter** [8]: This adaptive spatial filter shrinks the wavelet coefficients in a multi-resolution representation according to the probability of the presence of a signal of interest conditioned on a local spatial activity indicator.
- **The BLS-GSM filter** [5]: This method extends filtering in the steerable pyramid domain based on Gaussian Scale Mixtures [9] by employing a two-level (coarse-to-fine) local adaptation to spatial image features.
- **The Bishrink1, Bishrink2 filters** [11]: This method applies a bivariate shrinkage of the wavelet coefficients using the interscale dependencies and the local spatial variance estimation. Two variants were provided corresponding to different noise estimation levels.
- **The SPERRIL filter** [10]: This is an image restoration method, where the regularization (denoising) part is done in the steerable pyramid domain employing the interscale (parent-child) relationships between the coefficients.

3. PSYCHO-VISUAL EXPERIMENT

3.1 Experimental Setup

A psycho-visual experiment for the assessment of image quality has been described in detail in [6] for images artificially degraded by noise and blur. We constructed our own experiment that was slightly bigger and focused on artefacts that are more subtle.

Three scenes (Barbara, Face and Hill) containing different kinds of information ranging from texture over fine details to uniform backgrounds, see Fig. 1, were used in the experiment. These images were degraded by additive zero mean white Gaussian noise with a

(a) Barbara (b) Face (c) Hill

Figure 1: The test images used in our new psycho-visual experiment.

standard deviation of $\sigma = 15$ and $\sigma = 35$. The images were sent to the authors of the filters mentioned above who were asked to denoise them blindly, i.e. without any information on the noise level. Only the denoised images coming from $\sigma = 15$ were retained for the experiment since they showed least artefacts in general.
Then, the original image, the noisy one $\sigma = 15$, together with the 8 denoised images were presented to 37 subjects, on the same calibrated display, under comparable lighting conditions, on a 512×512 resolution. An example of the test images can be seen in Fig. 2 for Barbara. By prior inspection of the images we determined three different questions of interest for the subjects to decide on:

- How blurry are the images?
- How much artefacts are left in the images?
- How good is the overall image quality?

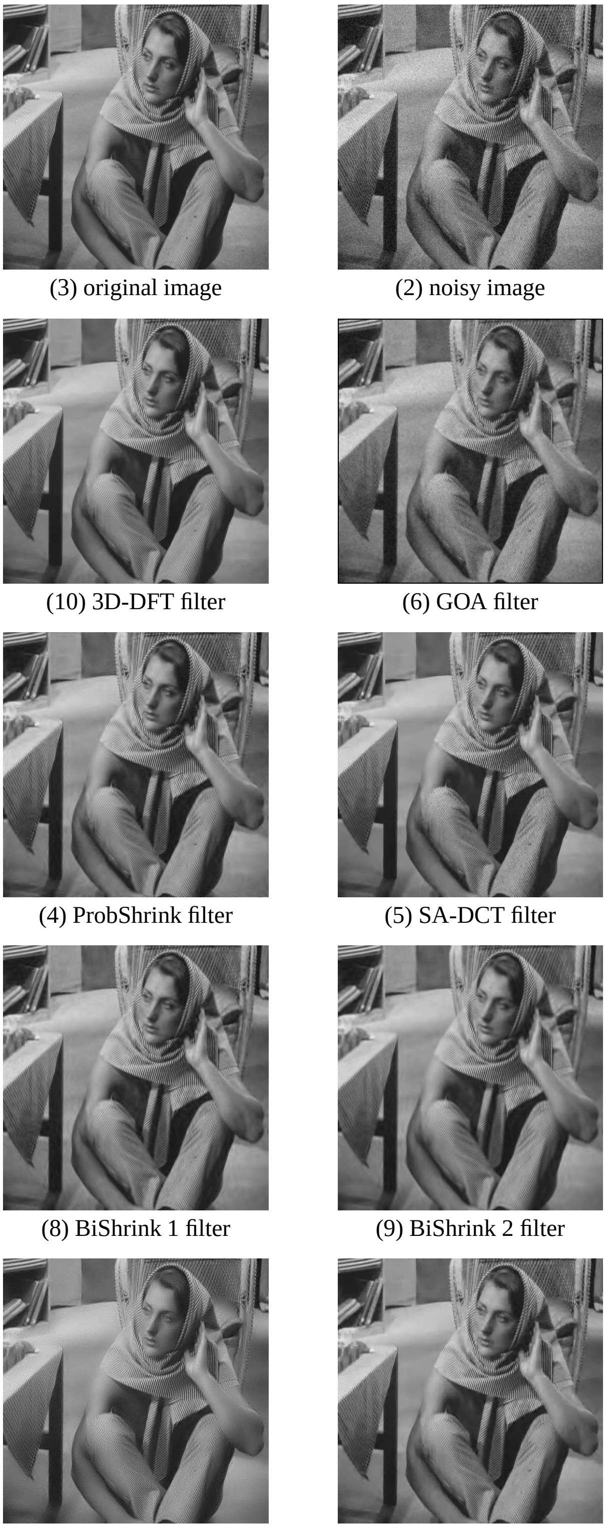
Questions one and two are expected to be in correlation to the third. First, 40 images were shown seperately one after the other, of which the first 10 are used to train the subject. The subjects were asked to score the attributes noisiness (artefacts), bluriness and overall image quality on a discrete scale from 0 to 5. Then 140 image couples were shown where image dissimilarity (how different one thinks the images are), on a scale from 0 to 5, and preference scores (which of the two images, left or right, one prefers in image quality), on a scale from -3 to 3, were asked.
In a follow-up experiment, 5 well-chosen triples of images were shown to 10 of the 37 subjects, who were then asked to retain the 2 best images and describe in words why they had retained them, this in order to better understand why one filter outperformed the other.

3.2 Multi-Dimensional Scaling

As mentioned in the introduction the MDS framework builds on the "principle of homogeneity of perception" combined with the concept of overall perceived image quality not being a 1-dimensional phenomenon. Therefore, out of different scores, such as the dissimilarity of two images in a pairwise experiment (two images shown simultaneously), the preference of two images again in a pairwise experiment and the attribute scores in a single-image experiment, a Multi-Dimensional geometric model, as can be seen in Fig. 3, is generated so that the inter-picture distance in the geometry corresponds linearly to the distances the different subjects have attached to them in the experiment. This geometry is obtained by a global recursive Maximum-Likelihood optimization combined with a χ^2 statistical model testing. The dimensionality of the geometry corresponds to the number of attributes needed and necessary to determine the optimal configuration. For technical details we refer to [7]. The XGms software tool available through the same book was used to perform these tests.

4. RESULTS

Fig. 3 shows the 2D-geometrical output configuration as optimized by the MDS framework from the combined results of the 37 subjects in the experiment. Each point corresponds to one of the filters shown in Fig. 2. The standard deviations on the positions are also plotted as the little ellipses. All stimulus positions were found statistically significant and similar configurations were obtained for



(3) original image (2) noisy image

(10) 3D-DFT filter (6) GOA filter

(4) ProbShrink filter (5) SA-DCT filter

(8) BiShrink 1 filter (9) BiShrink 2 filter

(7) SPERRIL filter (1) BLS-GSM filter

Figure 2: The test images for Barbara as presented in our own psycho-visual experiment.

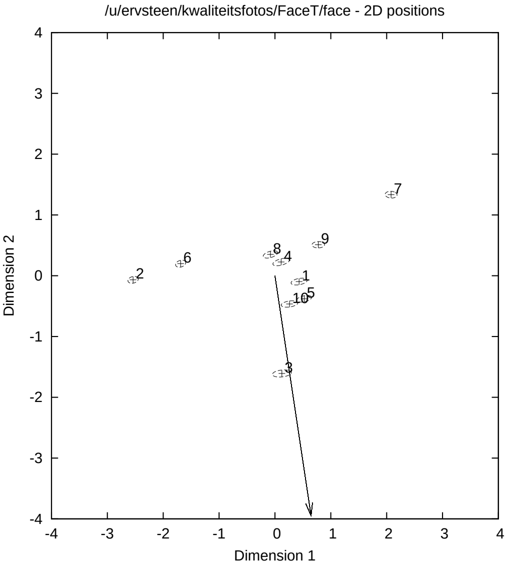


Figure 3: This figure shows the 2D-geometrical output of the MDS framework. The arrow points out the direction of the overall perceived image quality.

all three scenes. The black arrow points out the direction through which the overall image quality should be measured. The orthogonal projection of all points on this axis gives us a relative ranking of the images.

In that way, one can easily see that the original image (3) comes out best, followed by the 3D-DFT (10) and SA-DCT (5) filter, BLS-GSM (1) filter, ProbShrink (4) filter and Bishrink 1 (8). The GOA filter (6) and SPERRIL filter (7) are ranked worst, even below the noisy image (2). The Bishrink 2 (9) filter is ranked equally as the noisy image (2). A relative ranking of the perceived quality can be found in Fig. 6, notice that these numbers have no specific absolute meaning, the original image (3) is not found to be 5 times better then 3D-DFT (10) for instance.

A connotation to these distance can be given through an interpretation of the attribute axes, as we will show further on. In Fig. 3 we can also see the 3D-DFT (10) and SA-DCT (5) filter, BLS-GSM (1) filter, ProbShrink (4) filter and Bishrink (8,9) filters seem to somehow cluster, meaning that although they differ in ranking, they do have a common ground in terms of perceived image quality.

As a matter of comparison we also plotted the PSNR-values, see Table 1. From this table we see that 3D-DFT (10) still performs best in terms of PSNR but now the BLS-GSM (1) filter comes in second in case of Barbara, followed by SA-DCT (5), if we look at the bottom the table we also see some changes. Next to that, we also notice a shift in ranking through the scenes, which we don't have in the experiment. Finally, if we would compare the SA-DCT (5) and BLS-GSM (1) filters, see Fig. 4 and 5 for Barbara, it is clear the SA-DCT filter has slightly less blur and artefacts left, although there is an inverse difference in PSNR.

Another way of proving the PSNR is not suited best, is by computing the inter-pictures distance matrix D_p from the 2D geometrical MDS configuration and comparing this to the distance matrix D_s calculated through the PSNR. If the PSNR should predict the configuration well, then these matrices should be equivalent.

The equivalence between two matrices can be computed by the Spearman's Rank Order Correlation coefficient. Let D_p be the inter-picture distance matrix and D_s be a $N \times N$ similarity measure matrix, N being the number of input images used. Let

$$R_{ps} = 1 - 6\frac{\sum_{i=2}^{N}\sum_{j=1}^{i-1}(Rank[d_p(i,j)] - Rank[d_s(i,j)])^2}{N_D(N_D^2-1)}$$

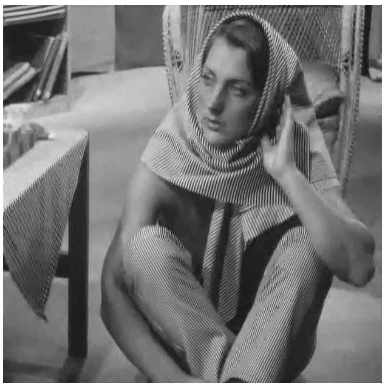


Figure 4: The result of the BLS-GSM filter which shows more artefacts (f.i. in the face) when compared to the result of the SA-DCT filter.

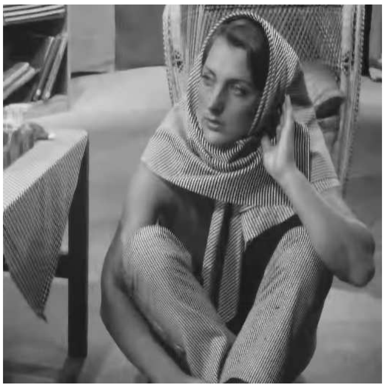


Figure 5: The result of the SA-DCT filter which shows less artefacts and better preserved texture than the result of the BLS-GSM filter.

where $Rank[d(i,j)]$ stands for the rank of matrix entry $d(i,j)$ which is a number between 1 and $N_D = N(N-1)/2$ when we order all matrix elements ascendingly. A derivative of the Spearman Rank Order Correlation coefficient D_{ps} is then given by

$$D_{ps} = \sqrt{1 - R_{ps}^2}.$$

One can easily see that this value ranges in the interval $[0,1]$ and that the smaller the value, the bigger the equivalence is between the matrices. For each of the three scenes we calculated this coefficient. The results vary from 0.586 for Barbara to 0.642 for Face to even 0.634 for Hill which means that the PSNR here does not coincide very well with visual perception.

	Face	**Barbara**	**Hill**
Noisy	24.65	24.62	24.65
3D-DFT	36.99	33.30	31.69
BLS-GSM	36.43	32.04	31.42
SPERRIL	31.09	28.99	28.50
GOA	21.25	23.45	22.26
ProbShrink	35.53	31.19	30.87
SA-DCT	36.82	31.38	31.54
Bishrink 1	36.47	31.20	31.01
Bishrink 2	36.27	29.76	29.95

Table 1: PSNR values (dB) for all three scenes and different filters

	original	3Swdft	SA-DCT	BLS-GSM	Genlik	Noisy	Bishrink2	GOA
original	/	*blur*	*blur*	*blur*	/	*noise*	/	*noise + blur*
3D-DFT	/	/	*detailinfo*	*blur + artefacts*	/	/	/	/
SA-DCT	/	/	/	*artefacts + details*	*blur + artefacts*	/	/	/
BLS-GSM	/	/	/	/	*blur + details*	/	/	/
Genlik	/	/	/	/	/	*blur*	*blur + noise*	/
Noisy	/	/	/	/	/	/	*blur*	*blur*
Bishrink2	/	/	/	/	/	/	/	*noise*
GOA	/	/	/	/	/	/	/	/

Table 2: This table shows the main attributes by which the filter in row i is chosen over the filter in column j, based on the follow-up experiment.

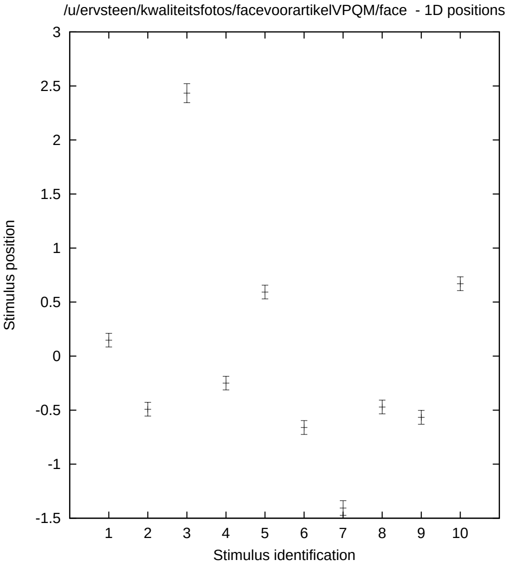
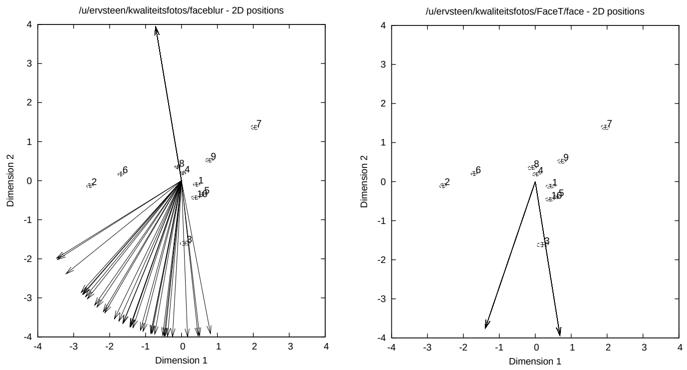
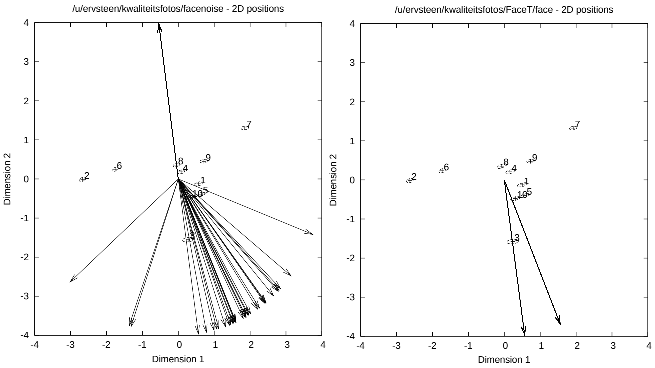


Figure 6: This figure shows the 1D-geometrical output of the MDS framework for Face. On the X-axis the different filters as numbered in Fig. 2, on the Y-axis the perceived quality.



(a) blur for different subjects (b) combined blur axis

Figure 7: On the left the 37 noise-axes are plotted for the individual subjects, on the right the optimized noise-axis. The quality axis is also plotted yet inversed on the left image for clarity reasons



(a) noise for different subjects (b) combined noise axis

Figure 8: On the left the 37 noise-axes are plotted for the individual subjects, on the right the optimized noise-axis. The quality axis is also plotted yet inversed on the left image for clarity reasons

We supposed our problem to be two-dimensional, meaning it is likely two attributes are adding up to the overall image quality, blur on the one hand and artefacts on the other. Since we also asked for bluriness and noisiness as extra attributes we can now try and predict those dimensions through the MDS framework.

However, Fig. 7 and 8 show us the directions of the individual axes of the 37 subjects for blur and noise, and the optimized common axis. Note that both axes should be read as descending blur and noise there where the quality is ascending (except for the case where we flipped the quality axis for clarity reasons). Since we assume the concept of “homegeneity of perception” we might wonder if the scope of the directions is not too broad to obtain the exact common direction, for sure in case of the noisiness. This means that different persons might score noisiness in different ways.

So instead of determining the exact direction of the blur and noise axis without further due, we performed a follow-up experiment as decribed in Section 3 for the Face image, selecting 10 subjects of which 5 had related attribute directions and 5 were considered outliers. The results from this follow-up experiment are shown in Table 2.

This table shows the main attributes taken into account by the different subjects in pointing out the actual difference between the images and should be interpreted as follows: the filter in row i outperforms the filter in column j, mainly based on table entry (i, j).

For instance, although the 3D-DFT (10) and SA-DCT-filter (5) are very close to one another as well in the 2D-configuration as in the 1D-projection in Fig. 3 and 6, their main difference lies in the amount of detail information left in 3D-DFT that is not present in SA-DCT. Also we can see that the noisy image is prefered above the Bishrink 2 (9) and GOA (6) filter, mainly because of the blur in the images. This means that although there is a lot of noise present, subjects tend to prefer the preservance of high frequency information and sharpness of edges in the images. This is in agreement with previous findings on the effect that noise has on images. Notice also that for time reasons not all possible triples were shown, but only those relevant to the quality ranking of Fig. 6.

From this table we can conclude that quality is highly related to the amount of remaining blur in the images. This is also partly justified by the findings of the MDS where as when we in fact determine one common blur axis from the different subjects, see Fig. 7, we obtain a direction strongly related to the quality axis as well as an ordening in bluriness corresponding to the follow-up experiment.
As for the noise-axis we see that what was called the noise-axis in fact really corresponds to a broader scope of artefacts, noise and detail information. Since most of the filters perform very well in terms of noise reduction, it is understandable the bigger part will be artefacts, which are harder to grasp in one dimension. Detail information can be related more to the amount of blur present.
This is why we choose to stick to Table 2 and the follow-up experiment to make further conclusions either than solely determining what we now will call the artefacts-axis through the MDS. Nevertheless, combining Fig. 7 (b) and Fig. 8 (b) we see that perceived overall quality is an (inverse) combination of blur and artefacts which is in accordance to Table 2, but this table provides us with extra information.
What we can do is relate the findings in Table 2 to the distances in the 2D-MDS configuration. In that way we can see that although the noisy image (2) and Bishrink2 (9) filter are very close in perceived overall quality, see Fig. 6, they are quite far apart in the 2D-geometry of Fig. 3. There is a relatively big difference in blur level, see Fig. 7 (b) as well as Table 2, a rather small difference in artefacts, see Fig. 8 (b) but still also a significant difference in high frequency noise which is not completely explained by the 2D-geometry. This might claim for an extra attribute-axis concerning the actual high-frequency (Gaussian) noise left, which is the purpose of future investigation.

5. CONCLUSIONS

The aim of this paper was to compare the perceived image quality of 7 state-of-the-art filters based on a psycho-visual experiment, leading to a ranking that is more true to human visual perception than instrumental images as the PSNR is. We were able to determine which of the filters related best to the original image, independent of the scene, namely the 3D-DFT and SA-DCT filters, followed by the BLS-GSM, ProbShrink and Bishrink 1 filter. We noticed from the MDS that, although the difference in quality, these five filters seem to cluster more or less, meaning that they show common grounds in terms of human visual perception.
In a follow-up experiment we were able to show why certain filters were found to outperform others and we could relate this to the findings of the MDS. Blurring and artefacts are shown to be the decision criteria, of which blur seems to carry the biggest load. As a preliminary overall conclusion we can say that for these type of filters, those that succeed in denoising images with minor blurring, even though this means leaving some of the noise present, while introducing minor artefacts, are considered perceptually best.

6. ACKNOWLEDGEMENTS

We would like to thank Prof. Dr. ir. Jean-Bernard Martens for the assistence on the XGms package and ir. Stefaan Lippens for the creation of the web-interface on which the psycho-visual experiment was performed http://telin.ugent.be/˜slippens/exp/ (in dutch). Especially, we would like to thank all authors taking part in this experiment.

REFERENCES

[1] Dabov K., Foi A., Katkovnik V. & Egiazarian K., “Image denoising with block-matching and 3D filtering”, *To appear in Image Processing: Algorithms and Systems V, 6064A-30, IST/SPIE Electronic Imaging*, 2006, San Jose, CA., 2006.

[2] Donoho D.L. & Johnstone I.M., “Ideal spatial adaptation by wavelet shrinkage”, *Biometrika*, vol. 81, no. 3, 1994, pp. 425-455.

[3] Escalante-Ramirez B., Martens J.B. & de Ridder H., “Multidimensional Characterization of the Perceptual Quality of Noise-reduced Computed Tomography Images”, *J. Visual Comm. Image Representation*, vol. 6, December 1995, pp. 317-334.

[4] Foi, A., Dabov K., Katkovnik V., & Egiazarian K., “Shape-Adaptive DCT for Denoising and Image Reconstruction”, *Proc. SPIE Electronic Imaging 2006, Image Processing: Algorithms and Systems V*, 6064A-18, San Jose, 2006.

[5] Guerrero-Colon J.A. & Portilla J., “Two-Level Adaptive Denoising Using Gaussian Scale Mixtures in Overcomplete Oriented Pyramids”, *Proceedings of IEEE ICIP conference*, Genoa, Italy, Sept 2005, pp 105-108.

[6] Kayagaddem V. & Martens J.B., “Perceptual Characterization of Images Degraded by Blur and Noise: experiments”, *Journal of Opt. Soc. Amer. A 13*, June 1996, pp. 1178-1188.

[7] Martens J.-B., “Image Technology Design”, *Springer*, 2003, Chapter 5.

[8] Pizurica A. & Philips W., “Estimating probability of presence of a signal of interest in multiresolution single- and multiband image denoising”, “A Joint Inter- and Intrascale Statistical Model for Bayesian Wavelet Based Image Denoising”, *IEEE Transactions on Image Processing*, in press.

[9] Portilla J., Strela V, Wainwright M & Simoncelli E.P., “Image Denoising using Scale Mixtures of Gaussians in the Wavelet Domain”, *IEEE Transactions on Image Processing*, vol. 12, no. 11, 2003, pp. 1338-1351.

[10] Rooms F., “Nonlinear Methods in Image Restoration Applied to Confocal Microscopy”, *Ph.D thesis*, 2005, Ghent University.

[11] Sendur L. & Selesnick I.W., “Bivariate Shrinkage With Local Variance Estimation”, *IEEE Signal Processing Letters*, vol. 9, no. 12, 2002, pp. 438-441.

[12] Sendur L. & Selesnick I.W., “Bivariate Shrinkage Functions for Wavelet-Based Denoising Exploiting Interscale Dependency”, *IEEE Trans. on Signal Processing*, vol 50, no. 11, 2002, pp. 2744-2756.

[13] Van De Ville D., Nachtegael M., Van der Weken D., Kerre E.E., Philips W., Lemahieu I., “Noise Reduction by Fuzzy Image Filtering”, *IEEE Transactions on Fuzzy Systems*, vol. 11, no. 4, 2003, pp. 429-436.

[14] Van der Weken D., Nachtegael M. & Kerre E.E., “Using Similarity Measures and Homogeneity for the Comparison of Images”, *Image and Vision Computing*, vol. 22 (9), 2004, pp. 695-702.

[15] Van der Weken D., Nachtegael M. & Kerre E.E., “Using Similarity Measures for Histogram Comparison”, *Lecture Notes in Artificial Intelligence*, vol. 2715, 2003, pp. 396-403.

[16] Van der Weken D., Nachtegael M. & Kerre E.E., “Using Similarity Measures for Histogram Comparison”, *Lecture Notes in Artificial Intelligence*, vol. 2715, 2003, pp. 396-403.