https://doi.org/10.2352/ISSN.2470-1173.2017.13.DPMI-083

BM3D-HVS: Content-adaptive denoising for improved visual quality

Karen Egiazarian[a,b], **Aram Danielyan**[b], **Nikolay Ponomarenko**[a,b], **Alessandro Foi**[a,b], **Oleg Ieremeiev**[c] , **Vladimir Lukin**[c]
[a] **Tampere University of Technology, 33101, Tampere, Finland**
[b] **Noiseless Imaging Ltd, 33720, Tampere, Finland**
[c] **National Aerospace University, 61070, Kharkov, Ukraine**

Abstract

We introduce a content-adaptive approach to image denoising where the filter design is based on mean opinion scores (MOSs) from preliminary experiments with volunteers who evaluated the quality of denoised image fragments. This allows to tune the filter parameters so to improve the perceptual quality of the output image, implicitly accounting for the peculiarities of the human visual system (HVS). A modification of the BM3D image denoising filter (Dabov et al., IEEE TIP, 2007), namely BM3D-HVS, is proposed based on this framework. We show that it yields a higher visual quality than the conventional BM3D. Further, we have also analyzed the MOSs against popular full-reference visual quality metrics such as SSIM (Wang et al., IEEE TIP, 2004), its extension FSIM (Zhang et al., IEEE TIP, 2011), and the no-reference IL-NIQE (Zhang et al., IEEE TIP, 2015) over each image fragment. Both the Spearman and the Kendall rank order correlation show that these metrics do not correspond well to the human perception. This calls for new visual quality metrics tailored for the benchmarking and optimization of image denoising methods.

Introduction

There has been intensive research on image denoising during the last few decades. Some argue that the practical limits of image denoising have now been reached. In fact, even though new methods and modifications of old methods keep being introduced, the relative progress in the quality of the denoised images seem to have become more and more insignificant: if the peak signal-to-noise ratio (PSNR) is used as a quality metric, the gap between the best denoising methods may be within only few tenths of decibel. Even according to established HVS-based quality metrics such as SSIM [1], this gap is becoming insignificant.

However, even the most advanced quality metrics when tested on specific databases of distorted images, fail to provide a satisfactory agreement with mean opinion scores (MOSs); for example, the state-of-the-art FSIMc [2] attains a Spearman rank order correlation coefficient (SROCC) for image database TID2013 [3] of only 0.85, whereas values of SROCC near unity are desired.

Due to HSV properties such as foveation and masking, perceptual quality assessment is inherently locally adaptive [1,4,5]. Image features and their statistical redundancy are also non-stationary, hence modern denoising filters employ various form of local adaptivity to the image content [6]. Thus, the perceptual optimization of a given filter may be approached by matching its local adaptivity to that of perceptual quality.

An additional layer of control on the local adaptivity of arbitrary filters can be obtained through the *content-adaptive filtering* [7], where a combination of elementary filters is applied to each image fragment in different proportions depending on the fragment's content. The elementary filters may be filters of different type or even a single filter using different parameter settings. The values of a local activity indicator (LAI) or several LAIs [8] are used to determine the adaptive combination or, more simply, the hard switching between different filters. It is important to choose LAIs, threshold(s), and filters that are well suited for the considered imagery [7,9].

In this paper, we propose a CAF design that is targeted at improving the perceived visual quality of the processed images. Instead of utilizing a specific visual quality metric, we employ MOS obtained from experiments with volunteers. Image fragments are denoised by different elementary filters and are shown to observers. We assume that the highest MOS for each considered fragment corresponds to the best suited elementary filter to be used by the designed CAF for filtering any such fragment. Hence, provided LAIs that can discriminate fragment classes corresponding to different elementary filters, we are able to improve the CAF perceptual quality without need of separately modeling either the relationship between image content and error visibility or the relationship between image content and action of the elementary filters.

The paper is structured as follows. First, we describe the proposed CAF framework. We then present the selection of LAI, features, elementary filters, and preparation of training image fragments. Next, we describe the setup of the experiment with human observers for obtaining MOS, followed by the statistical analysis of the collected results. We briefly analyze also how well several visual quality metric can predict the MOS. Finally, we present the designed CAF with hard switching, its denoising results, and conclusions.

Proposed CAF Design Framework

Let us represent a total of Q noisy image fragments with respect to a K-dimensional feature space with coordinates (features) $L_1,\dots,L_K$. Each noisy fragment is processed by D elementary filters Φ_d, $d=1,\dots,D$, yielding $Q\times D$ filtered fragments. These filtered fragments are shown to a, possibly large, group of human observers for perceptual evaluation. In particular, for each of the Q noisy fragments, each observer ranks the corresponding D filtered fragments based on the relative visual quality. This yields MOS values that indicate which filter is preferred for noisy fragments at specific $L_1,\dots,L_K$ coordinates. By regression of d on $L_1,\dots,L_K$, we can associate a preferred filter to each position in

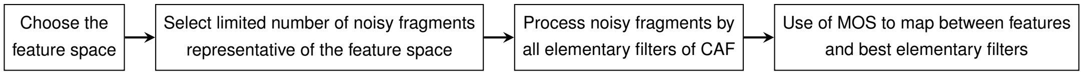

Choose the feature space
Select limited number of noisy fragments representative of the feature space
Process noisy fragments by all elementary filters of CAF
Use of MOS to map between features and best elementary filters

Figure 1. *Block diagram of the proposed framework*

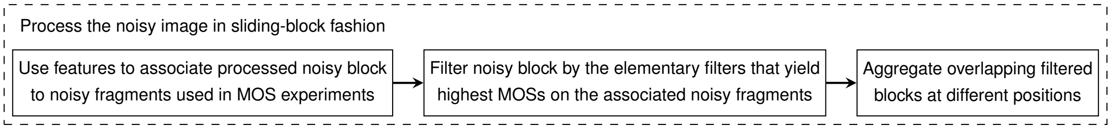

Process the noisy image in sliding-block fashion
Use features to associate processed noisy block to noisy fragments used in MOS experiments
Filter noisy block by the elementary filters that yield highest MOSs on the associated noisy fragments
Aggregate overlapping filtered blocks at different positions

Figure 2. *Block diagram of content-adaptive filter (CAF).*

the feature space as $\Phi_{d(L_1,\ldots,L_K)}$. Hence, a CAF is obtained by using the features $L_1,\ldots,L_K$ as LAIs.

The proposed general framework can be implemented in many ways. For example, the feature space may be partitioned, which is equivalent to a classification of the image fragments with respect to the feature values. Further, the feature-space selection and classification may be performed indirectly, e.g., through a neural network. The same trained network may be used for selecting representative noisy fragments for the subjective experiments, as well as employed by the CAF to classify image blocks and thus select the desired elementary filter. The CAF can be hard switching between the elementary filters, or soft switching (output of elementary filters are combined with weights dependent on the probability that the processed block belongs to the class for which the elementary filter is the preferred one). Furthermore, the MOSs can be based on either full-reference or no-reference subjective quality assessment.

Figures 1 and 2 illustrate the framework and the CAF.

In what follows, we present an instance of the proposed framework and CAF designed based on the BM3D filter [10].

Subjective experiments

As a representative set of noisy natural images, we consider grayscale versions of the Kodak images [11] cropped to 512×384 pixels as in the TID2013 database [3] and corrupted by additive white Gaussian noise (AWGN) with $\sigma^2 = 400$.

We consider a bivariate feature space ($K = 2$), defining the two LAIs as

$$L_1(X) = \frac{\sigma_X^2}{\sigma^2}, \qquad L_2(X) = \frac{1}{128\sigma^2}\sum_{i=1}^{8}\sum_{j=1}^{8}(X_{ij}-B_{ij})^2, \quad (1)$$

where X is a 8×8-pixel noisy block, σ_X^2 is the sample variance of X, σ^2 is the variance of the AWGN noise, and B is the most similar block to X in ℓ^2 sense. The noise variance σ^2 is assumed to be known or previously estimated. L_1 is used to characterize a local energy of the patch, and L_2 characterizes the level of dissimilarity of a given patch with respect to other nearby patches. The choice of these LAIs was due to the fact that BM3D exploits sparsity which results from both local smoothness and nonlocal self-similarity of an image.

Such small blocks are too small for the subjective experiments, for which we instead use 128×128-pixel test fragments extracted from the database images. To meet protocol recommendations on subjective assessment [12], we limit the duration of each experiment to 30 minutes and consider a total of $Q = 50$ noisy image fragments. In order to choose a limited set of fragments that is representative of the feature space defined by the L_1 and L_2 LAIs, we resorted to a clustering procedure which gave preference to fragments composed by a substantial majority of blocks that are well concentrated on the L_1, L_2 plane. The 50 fragments selected for our study are shown in Figure 3.

As the bank of elementary filters, we use the BM3D filter with $D = 8$ different values of hard threshold: 1.5σ, 1.9σ, 2σ, 2.5σ, 2.7σ, 2.9σ, 3.1σ, 3.5σ. The extremes of this range correspond to a very conservative preservation of details and edges with significant residual noise and to a very aggressive suppression of noise and significant smoothing of image content. Thus, for each of the 50 noisy fragments, we have a set of 8 filtered fragments to be evaluated in terms of visual quality.

In this work, we consider no-reference MOSs. During the subjective experiment, the participant is presented each of the 50 sets of filtered fragments in a random order. As can be seen in Figure 4), for each set, the noisy fragment is displayed at the center of the window, surrounded by the eight filtered fragments in random order. The participant has to select the filtered fragment with the best visual appearance. The selected fragment is then removed and the participant has to select the next best among the remaining ones. This is repeated until only the central noisy fragment is left. Thereafter, a new set of fragments is displayed. Auxiliary information, such as set presentation order, participant name, and total duration, is saved for each participant. Noise-free reference images were not shown to the participants.

A total of 125 volunteers participated to the perceptual experiments; they had no specific training or experience in image processing. The obtained data was analyzed and processed to detect and remove abnormal results according to methodology described in [3, 13]. For each fragment, MOS were measured robustly as trimmed mean, where individual scores considered abnormal were discarded in a two-stage sieving. Specifically, we first discarded all scores deviating from the median score more than 3/0.6745 times the median absolute deviation (MAD), as well as all scores from volunteers whose percentage of thus discarded scores exceeded 12%. Then, we rejected the remaining scores deviating from the mean score more than 2.33 times the standard deviation, and again all scores from the remaining volunteers whose percentage of thus discarded scores exceeded 12%. In this way, 33 subjects were discarded altogether, and 1646 scores from the remaining 98 valid subjects were also rejected as abnormal. Only the remaining 35154 scores (out of a total of 50000) were considered valid for the calculation of the MOS.

Table 1 gives an overview of the subjective experiments.

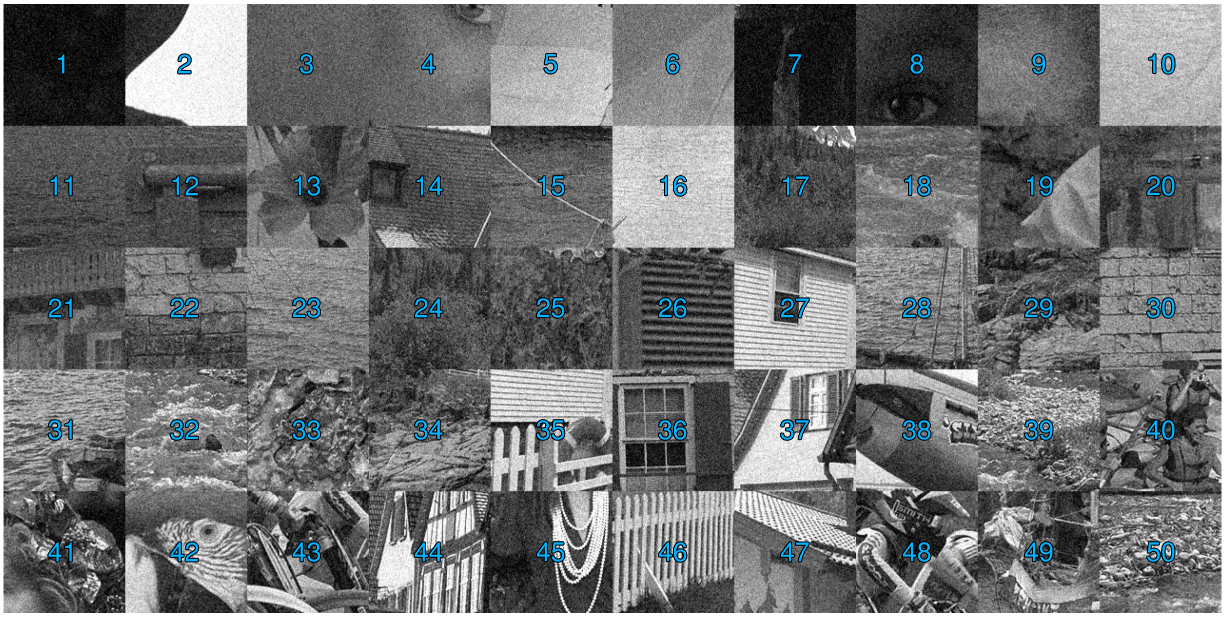
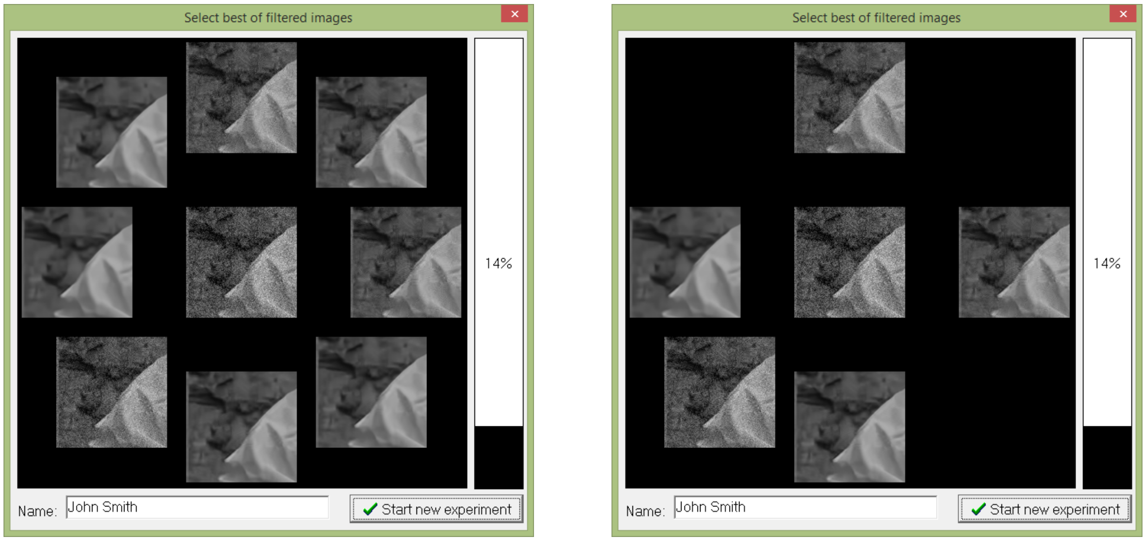


Figure 3. *The fifty 128×128 noisy image fragments used for generating the test sets for the subjective experiments. The blue numbers indicate the fragment number.*


Select best of filtered images
14%
Start new experiment
Select best of filtered images
14%
Start new experiment

Name: John Smith

Name: John Smith

Figure 4. *Screenshots of the user interface for assessing the quality of the fragments filtered by elementary filters: selection of the best filtered fragment at the beginning of the ranking (left); after the three best filtered fragments had been selected (right).*

Table 1. Overview of the subjective experiments

Number of noisy fragments	50
Number of elementary filters	8
Number of filtered fragments	400 = 50×8
Number of participants	125
Methodology of visual quality evaluation	Selection sort
Range of individual scores	0 (worst), . . . , 7 (best)
Number of individual scores	50000 = 125×50×8
Number of valid individual scores	35154

Results

Figure 5 reports the MOSs for each of the 8 elementary filters on each the 50 noisy fragments, i.e. the MOSs for each of the 400 filtered fragments. For any such filtered fragments, the MOS is the sample mean of the valid individual scores given by at most 92 valid subjects (87.9 on average). By dividing the sample standard deviation of the individual scores on each filtered fragment by the square root of the number of these scores, we obtain the standard deviations of the MOSs. In this way we obtain a confidence interval for each MOS, as illustrated in the figure. The MOS confidence intervals for the best performing filters on each fragment are rather wide and often intersect. This means that decision based on the largest MOS alone may not be perfect due to its noisy nature.

Table 2 presents Spearman and Kendall rank order correlation coefficients between MOS and a few good visual quality

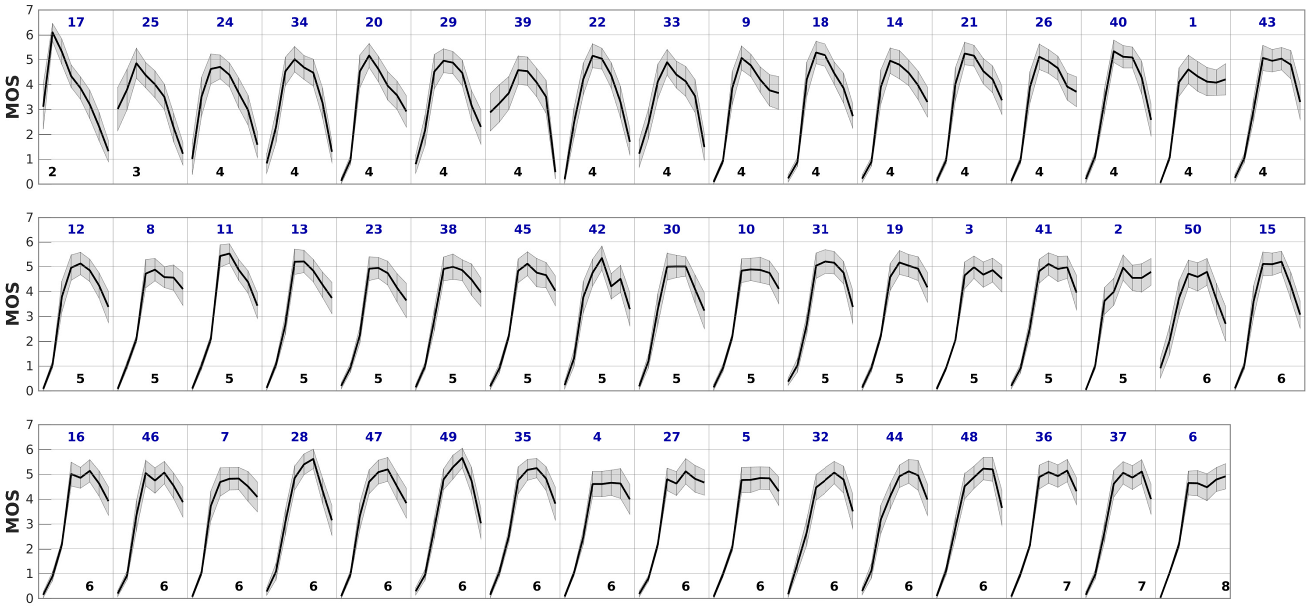


Figure 5. *Mean opinion scores (MOSs) for each of the 8 elementary filters on each the 50 noisy fragments, i.e. the MOSs for each of the 400 filtered fragments. MOSs are plotted in groups of 8 for each noisy fragment, which is indicated by the blue number on top of each subplot. The shaded gray area visualizes the* three-sigma *confidence interval for each MOS value. Noisy fragments are sorted according to the index of the corresponding processed fragment having highest MOS, indicated by the black number at the bottom of each subplot.*

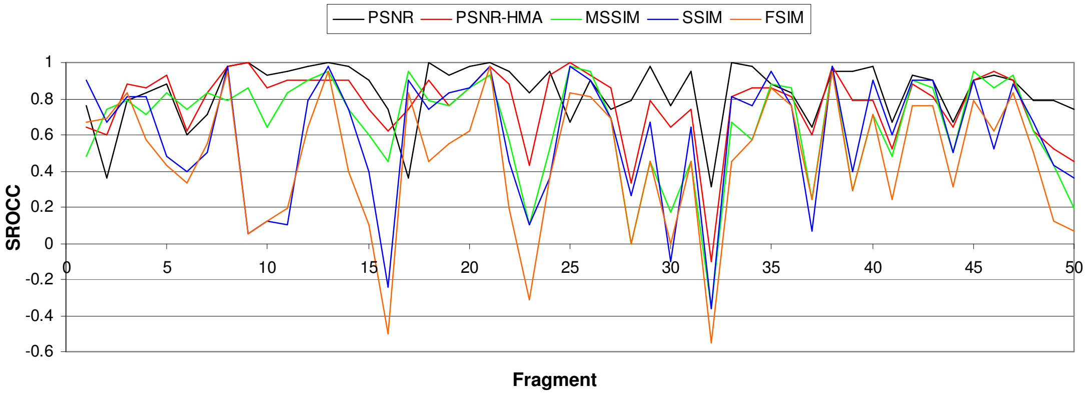


Figure 6. *Spearman rank-order correlation coefficients (SROCC) for the considered metrics and MOS, separately over each of the 50 groups of 8 filtered fragments.*

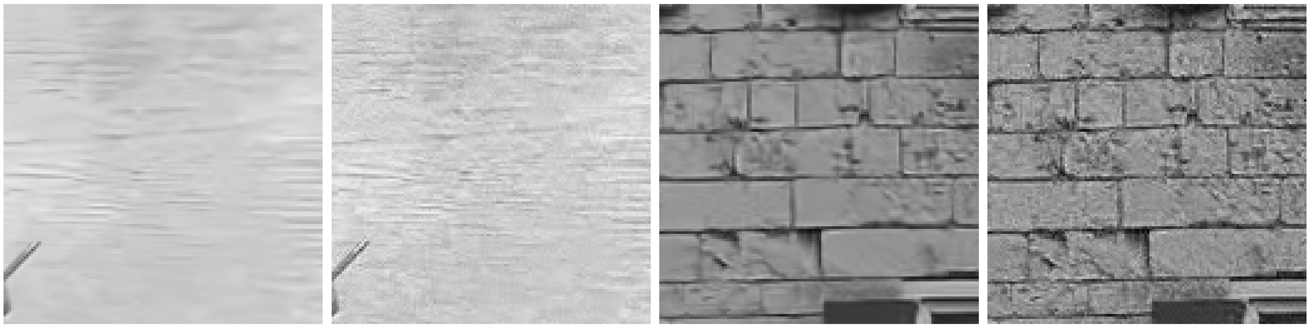
Table 2. Spearman and Kendall rank order correlation coefficients between MOS and metrics

Metric	Spearman correlation	Kendall correlation
PSNR	0.84	0.74
PSNR-HMA [16]	0.77	0.65
MSSIM [17]	0.65	0.55
SSIM [1]	0.59	0.51
NRSM [15]	0.52	0.44
FSIM [2]	0.46	0.40
IL-NIQE [14]	0.048	0.08

metrics over our test image sets. With the exception of the no-reference IL-NIQE [14] and NRSM [15], these are full-reference metrics which can be computed leveraging the noise-free fragments. No metric achieves satisfactory correlation coefficient values. Strikingly, the largest correlation coefficients are given by the PSNR, which does not take into account any peculiarity of the HVS. Even though FSIM provided one of the best results for the TID2013 [3] and LIVE [18] image databases, its performance is not satisfactory here. The plots in Figure 6 allow identifying the most problematic sets for the considered metrics. SROCC values on some sets (e.g., #32) can be even negative.

Figure 7 shows two examples of weak correspondence of SSIM to obtained MOS.

We argue that these general-purpose metrics are not suited for assessing the quality of denoised images because they are designed to address perception of the bias/variance trade-off on much larger ranges, such as considered in the above mentioned databases [3, 18].



a. *Fragment #16, threshold 2.9 MOS=5.14, SSIM=0.402*

b. *Fragment #16, threshold 2.0 MOS=2.17, SSIM=0.550*

c. *Fragment #30, threshold 2.9 MOS=5.01, SSIM=0.677*

d. *Fragment #30, threshold 2.0 MOS=3.28, SSIM=0.773*

Figure 7. *Two examples of weak correlation between SSIM and MOS*

Content-adaptive image denoising

As a result of the subjective experiments, for each of the 50 noisy fragments we know which filter, among the eight variants, provides the highest MOS; we form a vector **b** of length 50 with values from 1 to 8 indicating the best filter for each fragment.

We implement the CAF as follows.

1. Compute outputs $\mathbf{y}^{(t)}$, $t = 1,\ldots,8$ of eight BM3D filters with the corresponding thresholds $\{1.5\sigma, 1.8\sigma, 2\sigma, 2.5\sigma, 2.7\sigma, 2.9\sigma, 3.1\sigma, 3.5\sigma\}$.

2. Calculate $L_1(X)$ and $L_2(X)$ values (1) for a sliding 8×8 block X in the noisy input image, resulting in 2D arrays Z_{L_1} and Z_{L_2}. Perform the same procedure for the 50 noisy fragments, resulting in the 2D arrays $V_{L_1}^{(k)}$ and $V_{L_2}^{(k)}$, $k = 1,\ldots,50$.

3. For a given 8×8 noisy block at position i and j (coordinates of the top-left corner pixel) form a vector $(Z_{L_1}(i,j), Z_{L_2}(i,j))$ and perform the following. For each of the $k = 1,\ldots,50$ noisy image fragments, calculate the number $\mu(k)$ of points $(V_{L_1}^{(k)}, V_{L_2}^{(k)})$ falling inside of the circle with radius T around $(Z_{L_1}(i,j), Z_{L_2}(i,j))$ within the L_1, L_2 plane, as illustrated in Figure 8. The value $T = 0.55$ is chosen empirically. Let k_1, k_2, k_3 be the indexes corresponding to the three largest values of $\mu(k)$, $k = 1,\ldots,50$. Then, $t_1 = b(k_1)$, $t_2 = b(k_2)$ and $t_3 = b(k_3)$ are the indexes of three best filters (among the 8 variants) for the given patch at the position (i,j). We define the *partial* weights $w_{i,j}$ as follows:

$$w_{i,j}(i:i+7, j:j+7, t_1) = \mu(k_1),$$
$$w_{i,j}(i:i+7, j:j+7, t_2) = \mu(k_2),$$
$$w_{i,j}(i:i+7, j:j+7, t_3) = \mu(k_3),$$

and $w_{i,j}$ is zero elsewhere.

4. The partial weights $\mathbf{w}_{i,j}$ are accumulated for all block positions, giving the CAF weights $\mathbf{w}$ as

$$\mathbf{w} = \sum_{i,j} \mathbf{w}_{i,j}\,.$$

5. The output $\mathbf{y}$ of the designed filter at the position (i,j) is computed as the linear combination of the outputs of $\mathbf{y}^{(t)}$, $t = 1,\ldots,8$,

$$\mathbf{y}(i,j) = \frac{\sum_{t=1}^{8} w(i,j,t)\mathbf{y}^{(t)}(i,j)}{\sum_{t=1}^{8} w(i,j,t)}\,.$$

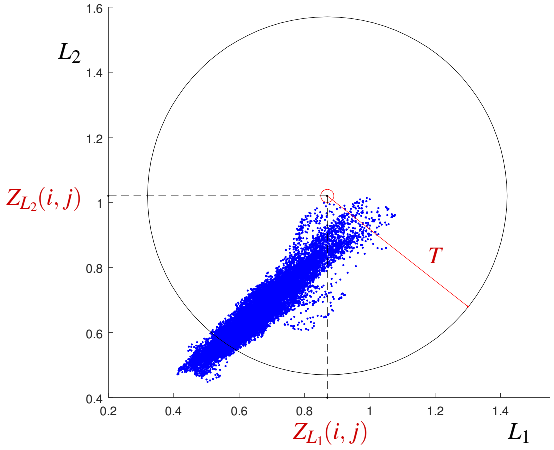
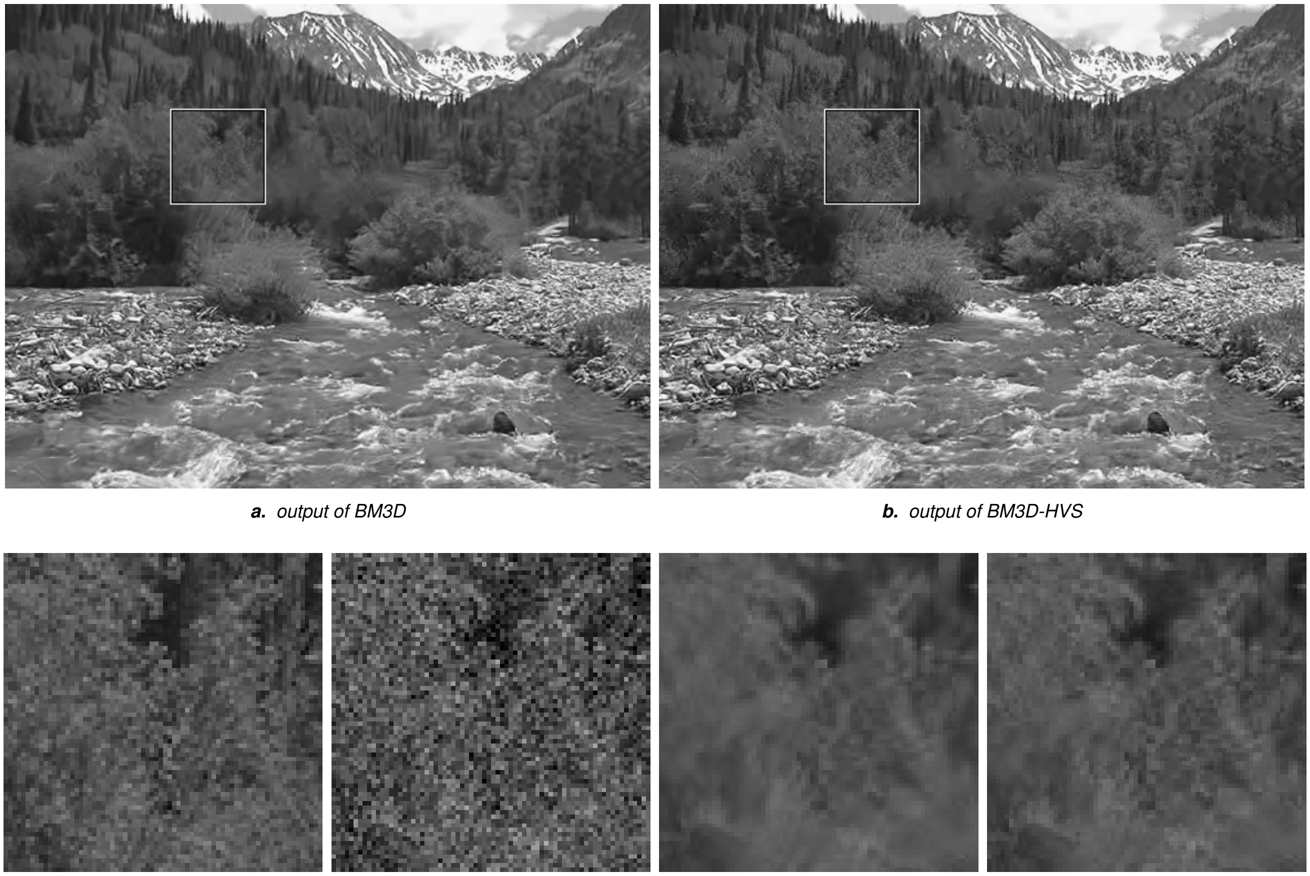


Figure 8. *Scatterplot of $(V_{L_1}^{(k)}, V_{L_2}^{(k)})$ (blue dots) for fragment #1; the disc of radius $T = 0.55$ is the area of calculation of $\mu(1)$ for a given $(Z_{L_1}(i,j), Z_{L_2}(i,j))$.*

Figure 9 shows an example of filtering the test image #13 from the TID2013 database by the standard BM3D filter and by BM3D-HVS obtained from the proposed CAF design. One can observe that the synthesized filter provides better details preservation than the standard filter add noise suppression in uniform areas has not changed. Output images in Figure 9 and four other such examples were shown, without providing background information, to 10 observers (undergraduate students) who unanimously identified the outputs of BM3D-HVS as having better visual quality than outputs of the conventional BM3D filter. Figures 10 and 11 provide a further illustration of the qualitative differences between BM3D and BM3D-HSV.

Conclusions

We have presented a content-adaptive approach to image denoising. We have used the BM3D filter as an example to validate our framework and performed a series of experiments with volunteers who have evaluated the quality of image fragments denoised by BM3D with different threshold values (elementary filters), collecting MOS values. We designed a CAF as a linear combination of the outputs of the elementary filters with spatially



a. *output of BM3D*

b. *output of BM3D-HVS*

c. *Fragment of the noise free image*

d. *Fragment of the noisy image*

e. *Magnified fragment of (a)*

f. *Magnified fragment of (b)*

Figure 9. *The results of filtering the test image #13 from the database TID2013 corrupted by AWGN with variance σ^2=400.*

adaptive weights based on obtained MOSs. This CAF provides improved visual quality of the output image, implicitly accounting for the peculiarities of the HVS through its MOS-based design. We further analyzed the correlation between the collected MOSs and several image quality metrics, reference-based as well as non-reference, finding a significant mismatch. Thus designing a quality metric appropriate for image denoising remains an open problem.

Acknowledgments

This work is partially supported by Academy of Finland, project no. 287150, 2015-2019.

References

[1] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[2] L. Zhang, L. Zhang, X. Mou, and D. Zhang, “FSIM: a feature similarity index for image quality assessment,” *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, 2011.

[3] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti *et al.*, “Image database TID2013: Peculiarities, results and perspectives,” *Signal Process.: Image Comm.*, vol. 30, pp. 57–77, 2015.

[4] J. M. Foley, “Human luminance pattern-vision mechanisms: masking experiments require a new model,” *JOSA A*, vol. 11, no. 6, pp. 1710–1719, 1994.

[5] J. A. Solomon, A. B. Watson, and A. Ahumada, “Visibility of dct basis functions: Effects of contrast masking,” in *Proc. Data Compr. Conf., 1994 (DCC’94)*. IEEE, 1994, pp. 361–370.

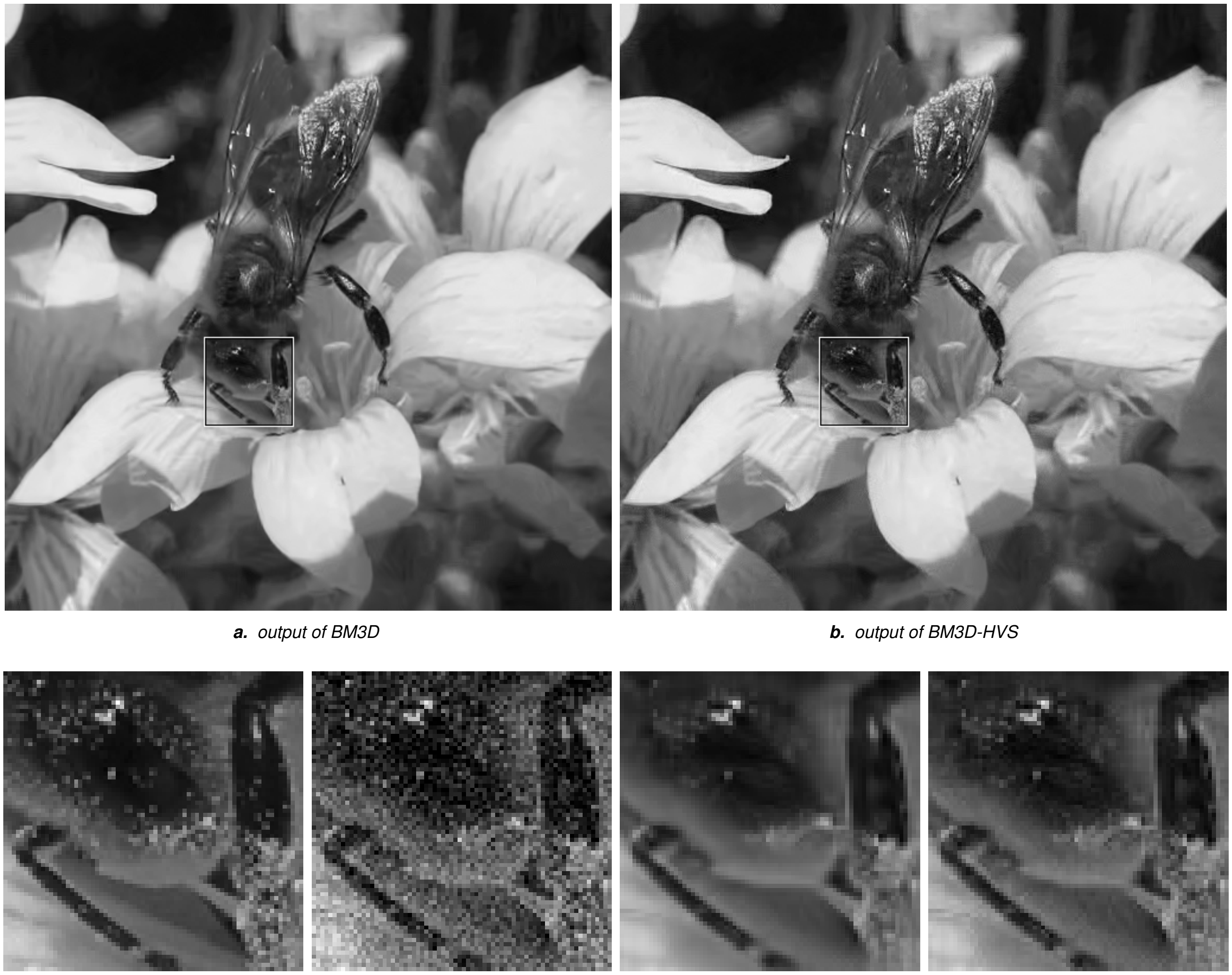
[6] V. Katkovnik, A. Foi, K. Egiazarian, and J. Astola, “From local kernel to nonlocal multiple-model image denoising,” *Int. J. Comput. Vision*, vol. 86, no. 1, pp. 1–32, 2010.

[7] V. P. Melnik, V. V. Lukin, A. A. Zelensky, J. T. Astola, and P. Kuosmanen, “Local activity indicators for hard-switching adaptive filtering of images with mixed noise,” *Optical Engineering*, vol. 40, no. 8, pp. 1441–1455, 2001.

[8] V. P. Melnik, V. V. Lukin, A. A. Zelensky, H. Huttunen, and J. T. Astola, “Nonlinear locally adaptive and iterative algorithms of image restoration,” *J. Electronic Imaging*, vol. 6, no. 4, pp. 439–452, 1997.

[9] N. N. Ponomarenko, V. V. Lukin, A. A. Zelensky, K. O. Egiazarian, and J. T. Astola, “Locally adaptive image filtering based on learning with clustering,” in *Electronic Imaging 2005*. International Society for Optics and Photonics, 2005, pp. 94–105.

[10] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, “Image denoising by sparse 3-D transform-domain collaborative filtering,” *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080–2095, 2007.



a. *output of BM3D*

b. *output of BM3D-HVS*

c. *Fragment of the noise free image*

d. *Fragment of the noisy image*

e. *Magnified fragment of (a)*

f. *Magnified fragment of (b)*

Figure 10. *The results of filtering the test image Fly corrupted by AWGN with variance σ^2=400.*

[11] “Kodak lossless true color image suite,” PhotoCD PCD0992. [Online]. Available: http://r0k.us/graphics/kodak/

[12] ITU Radiocommunication Assembly, *Methodology for the subjective assessment of the quality of television pictures*. International Telecommunication Union, 2003.

[13] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, “A statistical evaluation of recent full reference image quality assessment algorithms,” *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, 2006.

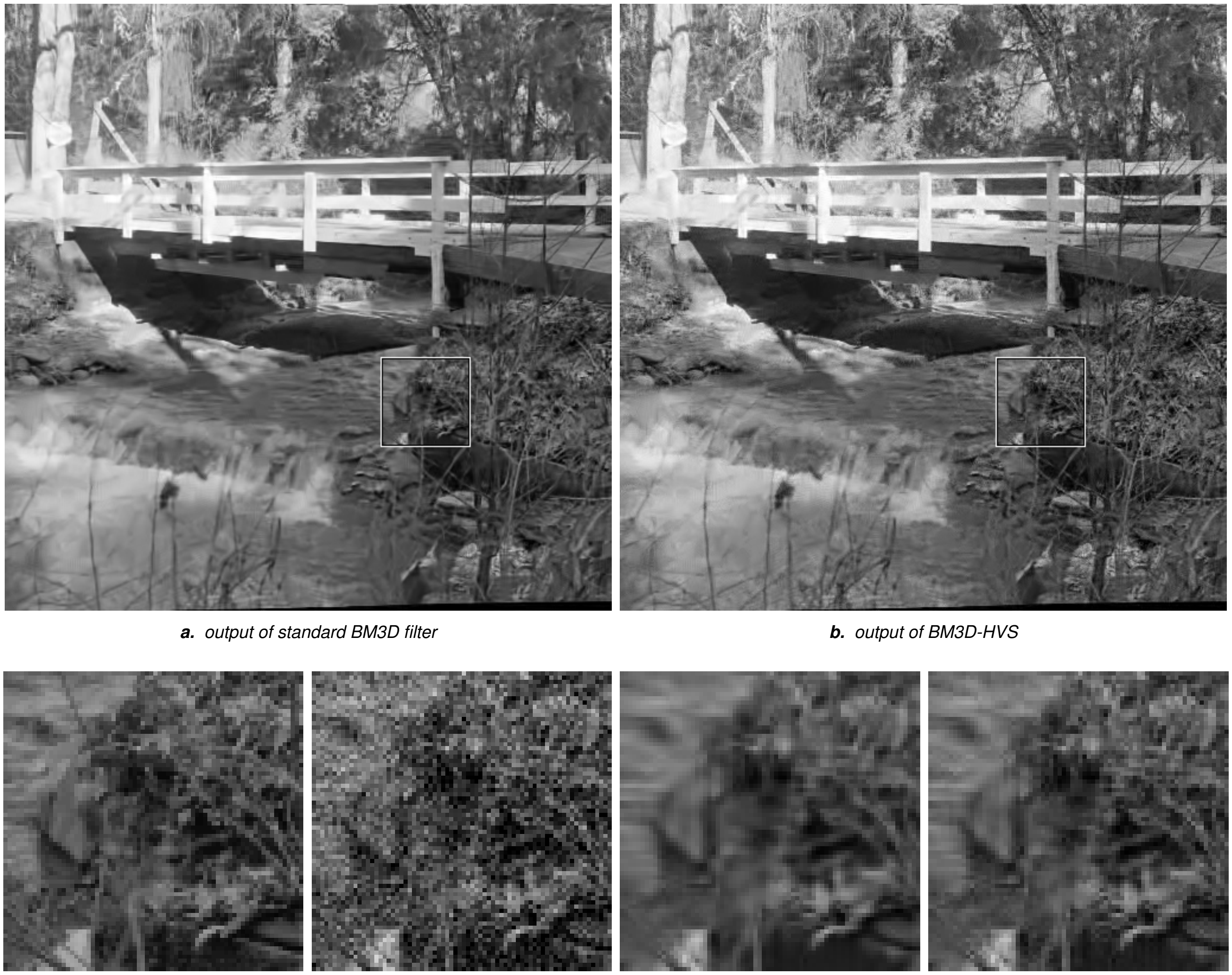
[14] L. Zhang, L. Zhang, and A. C. Bovik, “A feature-enriched completely blind image quality evaluator,” *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2579–2591, 2015.

[15] N. N. Ponomarenko, V. V. Lukin, O. I. Eremeev, K. O. Egiazarian, and J. T. Astola, “Sharpness metric for no-reference image visual quality assessment,” in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2012, pp. 829 519–829 519.

[16] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, and M. Carli, “Modified image visual quality metrics for contrast change and mean shift accounting,” in *CAD Systems in Microelectronics (CADSM), 2011 11th International Conference The Experience of Designing and Application of*. IEEE, 2011, pp. 305–311.

[17] Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multiscale structural similarity for image quality assessment,” in *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*, vol. 2. IEEE, 2003, pp. 1398–1402.

[18] H. R. Sheikh, Z. Wang, A. C. Bovik, and L. Cormack, “Image and video quality assessment research at live,” http://live.ece.utexas.edu/research/quality/, 2003, [Online; accessed January 17, 2017].



***a.** output of standard BM3D filter*

***b.** output of BM3D-HVS*

***c.** Fragment of the noise free image*

***d.** Fragment of the noisy image*

***e.** Magnified fragment of (a)*

***f.** Magnified fragment of (b)*

Figure 11. *The results of filtering the test image Bridge corrupted by AWGN with variance σ^2=400.*