Extracting Compact Information from Image Benchmarking Tools: The SAR Despeckling Case

Gerardo Di Martino[1], Giovanni Pecoraro[2], Giovanni Poggi[1], Daniele Riccio[1], and Luisa Verdoliva[1]

[1] DIETI, University Federico II of Naples, Naples (I)
[2] Accademia Aeronautica, Pozzuoli (I)
{firstname.lastname}@unina.it

Abstract. Image databases and benchmarks are precious tools to assess the quality of competing algorithms and to fine tune their parameters. In some cases, however, *quality* cannot be captured by a single measure, and several of them, providing typically contrasting indications, must be computed and analyzed. This is certainly the case for the SAR despeckling field, also because of the lack of clean reference images, which forces one to compute the measures of interest on simple canonical scenes. We present here the first results of an ongoing work aimed at selecting a suitable combination of benchmark measures to assess competing SAR despeckling techniques and rank them. The full validation of the proposed methodology will require the involvement of a reasonable number of expert photo-interpreters for a large-scale experimental campaign. Here, we present only a sample experiment to provide some insight about the approach.

Keywords: Reproducible research, image benchmarking, SAR despeckling.

1 Introduction

Modern research in experimental sciences relies heavily on the concept of reproducibility, which requires researchers to put colleagues in the condition to replicate, and hence validate, their experiments, by thoroughly describing them and providing reference data.

Although this good practice has been long established in many fields of science, it is not so widespread in signal and image processing [1], especially when applied to remote sensing. A large number of scientific papers present experimental results obtained on test images that are not available to fellow researchers, with algorithms only summarily described and whose source/executable code is also unavailable. Commercial or intellectual property issues justify, sometimes, this behavior, but a drift towards fully reproducible research is obviously necessary.

The diffusion of reliable and thorough image databases and benchmarks is a keystone in this path. Databases allow experimenting with the same data used by others and guarantee conformity of data to standard requirements. Then, given common data, and the executable codes of competing techniques, a benchmarking tool allows one to compare a large number of techniques and choose the one that best fits the application at hand. In addition, it allows one to fine-tune the parameters of an algorithm to obtain the best possible performance.

A. Petrosino (Ed.): ICIAP 2013, Part II, LNCS 8157, pp. 11–20, 2013.

The above picture, however, is overly simplified since, even with plenty of benchmark data, it is not always obvious how to choose the best technique, because there is rarely a single performance metric that thoroughly qualifies quality. In image denoising, for example, the squared error (hence SNR, PSNR, and the likes) has long been the performance metric of choice. However, it is well known that for a human being, quality is not always well correlated with squared error, and there has been an intense search for alternative measures, with the structural similarity (SSIM) [2] now considered as a valid alternative. Although, most of the times, SNR and SSIM provide similar indications, the latter is more sensitive to various forms of image impairment which are relevant for perceived quality. Things become much more complex when the problem under investigation calls naturally for many different quality metrics. In edge detection, two measures are necessary, related to false and missed edges, which can be still managed through simple performance curves as in the well-known Berkeley database [3]. To measure image segmentation performance, however, a large number of indicators can be legitimately considered, as shown for example in the Prague remote sensing segmentation benchmark [4,5]. Given such a wealth of measures, rarely pointing towards a clear winner, the problem becomes how to extract useful indications from them. This problem is even more pressing if one sees the benchmarking tool as instrumental to fine-tuning a given algorithm. Image processing can rarely avoid the setting of thresholds and other parameters, usually selected with a grain of salt by the designer or the user. A good benchmarking tool could help selecting such parameters in a more objective and robust way.

Extracting compact and reliable performance information from a large set of contrasting indicators is a quite general problem, investigated in such different areas as economics [6] and bibliometrics [7], and is certainly relevant for image processing as well. Here, we focus on SAR (Synthetic Aperture Radar) image denoising, which fits very well the above description. In this field, direct objective measurements of quality are not possible, and one must rely on indirect indicators which account for such diverse items as noise suppression, edge preservation, radiometric and spatial resolution preservation. Typically, techniques that perform well under some points of view, do not under some others, and hence it is not easy to establish a meaningful ranking, nor to guide the selection of optimal parameters. We propose a methodology to define a good combination of indicators or, at least, to choose the best among some proposed combinations, relying on the comparison between the behaviour of the combination and the mean opinion score of expert photo-interpreters.

In next Section we provide some detail on the SAR despeckling problem and on the related benchmarking tool used in this work. Section III describes the proposed methodology for selecting a suitable combination of indicators. Finally Section IV provides a sample experimental result and outlines future work.

2 The SAR Despeckling Benchmark

Synthetic Aperture Radar (SAR) sensors are valuable sources of remote-sensing imagery. Mounted on satellites or planes, they can collect images of the surface irrespective of illumination and cloud coverage, with a spatial resolution that, thanks to the

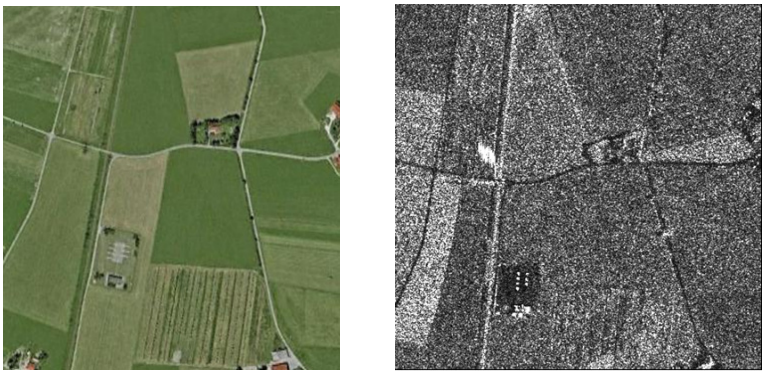


Fig. 1. Optical vs SAR images of the same region (not the same time instant, not co-registered)

synthetic aperture principle, can go below 1 meter. Because of the imaging modality, however, SAR images are affected by a severe noise, called speckle, which, under suitable hypotheses, is well modeled as multiplicative and spatially uncorrelated.

Fig.1 compares an optical image with its SAR counterpart, the latter exhibiting the peculiar "salt-and-pepper" appearance due to speckle noise. Regions that look homogeneous in the optical image are not anymore in the SAR image and, what is worse, fine details easily spotted in the optical case are hardly recognizable with SAR. To allow for an easier interpretation and automatic processing of SAR images, for example by means of suitable object-oriented segmentation and representation tools [8], a large number of despeckling filters have been proposed in the literature, which are necessarily different from conventional denoising filters because of the different type of noise involved and also for the different properties of SAR images w.r.t. optical ones.

Comparing the performance of such techniques is a challenging task by itself, since no such thing as a "clean" SAR image exists to compute full-reference measures. In fact, speckle is an inherent and hence unavoidable feature of SAR images, which can be reduced only by spatial averaging, renouncing full spatial resolution, or by averaging multiple instances of the same scene taken at close instants, which is very difficult because of technological limitations. Therefore, to assess a despeckling technique one has to follow indirect paths, such as using *ad hoc* no-reference measures which account only for some of the phenomena of interest, or simulating SAR images based on optical images, a shaky practice given the deep differences between these types of sources, and leaving the final word, in any case, to visual inspection by experts.

A new approach to SAR despeckling assessment was recently proposed in [9], based on the physical level, hence realistic, simulation of synthetic SAR images by means of the SARAS simulator [10]. Given the electromagnetic and geometrical properties of the surface, and the SAR system parameters, the SARAS generates the corresponding SAR image, which is deterministic but for the speckle. By averaging an arbitrary number of instances of such scenes, differing only for speckle content, we are able to provide a legitimate speckle-free image to compute all necessary full-reference measures. Of course, since the characteristics of the scene must be defined by the designer, only

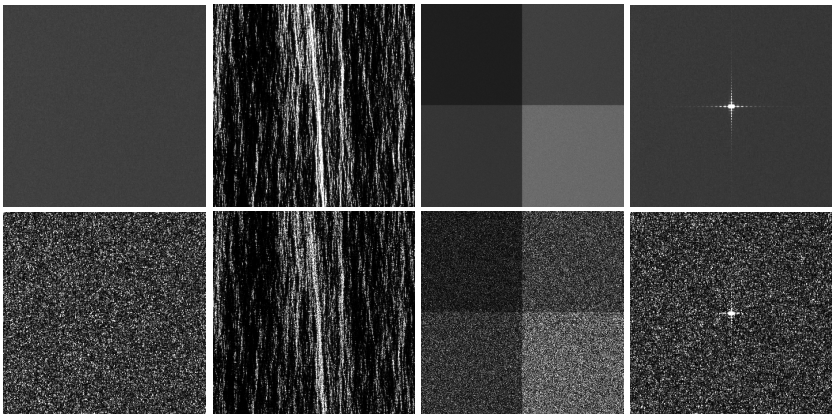


Fig. 2. Canonical scenes for the benchmark: Homogeneous, DEM, Squares, Corner. Upper row: 512-look reference images, bottom row: 1-look test images.

simple scenes can be generated, composed by a relatively small number of regions, each one characterized by a few parameters.

Therefore, the proposed benchmarking tool considers a small set of canonical scenes, some of which are shown in Fig.2, a homogeneous flat region, a textured one, a mosaic of several regions with edges, a corner reflector on flat background, designed so as to measure the features most relevant for SAR despeckling quality. In particular, a good despeckling filter should possess all the following characteristics:

- strong speckle reduction in homogeneous areas;
- scene feature preservation (like textures, edges, point targets);
- radiometric preservation;
- absence of artifacts.

Accordingly, the benchmarking tool considers the following measures, computed with respect to the 512-look reference, and described here in qualitative terms (the reader is referred to [9] for more detail)

- MoI/H: Mean of Intensity, bias on mean value of the filtered homogeneous scene (Homog);
- ENL: Equivalent Number of Looks, flatness of Homog;
- DG/H: Despeckling Gain, SNR gain due to despeckling on Homog.;
- DG/D: Despeckling Gain, SNR gain due to despeckling on DEM;
- C_x: Coefficient of variation, activity of DEM;
- ES: Edge Smearing, edge profile preservation on Squares;
- FOM: Figure Of Merit, edge location preservation on Squares;
- C_{BG}: Contrast to BackGround, radiometric fidelity on Corner.

An example table of results is reported in Tab.1 for a few state-of-the-art despeckling algorithms: enhanced Lee filter [11], based on adaptive spatial filtering, PPB [12], SAR-BM3D [13,14], and FANS [15], based on the nonlocal approach, the latter being a fast

adaptive version of SAR-BM3D. In the first two rows we report the ideal value, obtained when the filtered image equals the 512-look reference, and the value obtained on the 1-look noisy image. We avoid any specific comment on results, being out the scope of this paper, observing only that the resulting figures are hardly comparable with one another and provide often contrasting indications of quality, based on which it is difficult or at least controversial to decide which technique performs best.

Table 1. Despeckling benchmark measures for some selected techniques

	MoI/H	ENL	DG/H	DG/D	C_x	ES	FOM	C_{BG}
512-look	1.000	515.57	∞	∞	2.40	0	0.993	36.56
1-look	0.987	0.99	0	0	3.55	0.105	0.792	36.54
Lee	**1.003**	40.07	15.23	1.90	2.86	0.392	0.797	**36.44**
PPB	1.005	135.54	20.18	3.63	2.71	0.334	0.837	33.92
SAR-BM3D	0.984	99.71	19.18	**5.19**	**2.45**	**0.222**	**0.847**	35.58
FANS	1.014	**147.19**	**20.25**	4.87	2.57	0.361	0.776	35.83

3 Methodology

Our ultimate goal is to learn how to combine the indicators obtained on the canonical scenes by competing despeckling techniques in order to predict their actual ability to correctly despeckle real-world SAR images. Lacking a mathematical model that relates the measures with the despeckling power, we need the largest possible set of experimental data, namely, a collection of observed benchmark measures with attached a "true" quality measure, so as to compute all desired correlations and find eventually a reliable prediction rule.

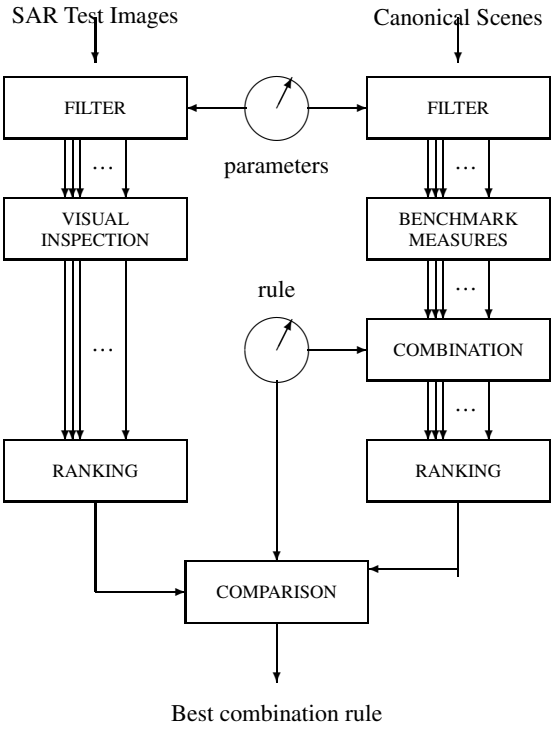
In practice we will come much shorter of this goal for several good reasons, first of all, the lack of a reliable and objective (computable) measure of quality for despeckled SAR images. Even if we had a clean reference of the considered scene, which we do not, a single measure like the MSE would be questionable, as explained in Section II. We can therefore consider two alternatives approaches:

1. use objective performance indicators of some subsequent automatic task carried out on the despeckled images;
2. collect the mean opinion score of a panel of expert photo-interpreters.

Both approaches have their drawbacks. The first one depends on the type of application considered (e.g., classification, target detection, etc.) and on the specific tool used: different applications and tools would certainly lead to different scores for the same technique, undermining the desired objectivity. This problem is always present, because image quality depends intrinsically on the intended application and even photo-interpreters have indeed their own points of view and expectations about SAR images.

Humans, however, are certainly better able to weight the various sources of impairment according to their relative importance expressing eventually a more balanced judgement. Therefore we follow this second approach, here, taking also advantage of the professional expertise of some of the Authors in SAR image analysis. However, humans cannot associate a meaningful numerical score with the images so we replace the absolute quality score with the ranking. Interpreters are asked to rank the despeckled images in order of quality by comparison, and the ranking will be eventually used as score.

Rather than using different techniques, in this preliminary work we use a single technique and change some relevant parameters that strongly impact on the performance. Specifically, we consider our FANS algorithm, based on the state-of-the-art SAR-BM3D technique. FANS runs about 10 times faster than SAR-BM3D. Apart from this detail, the major modification w.r.t. the original algorithm consists in the use of a classification step: each image patch is classified as either active (texture, edge, permanent scatterer) or flat by comparing a suitable statistic with a threshold selected by the user. Active blocks are filtered considering a large analysis window, while flat blocks use a smaller one. By tuning the threshold and the size of the large and small search areas significant differences in the performance are observed, suggesting that a correct tuning of such parameters is crucial for the overall performance.



The methodology of analysis is summarized in the above scheme. FANS parameters are changed in a wide range, selected in advance, thus modifying the algorithm

behavior. The same filter, with the same parameter, is used to filter, on two parallel paths, both N_{SAR} real-world SAR images and the benchmark canonical scenes. By varying the parameters, a number of filtered SAR images are collected (represented by the multiple lines departing from the filter) and given to N_{PI} photo-interpreters for quality ranking. For a given selection of the parameters, we thus obtains $N_{\mathrm{SAR}} \times N_{\mathrm{PI}}$ ranking scores which are simply averaged to obtain a mean opinion score (MOS) and hence the final ranking. On the other path, the filter is used for the canonical scenes and for each choice of the parameters we obtain N_{BI} benchmark indicators which must be combined according to some suitable rule to provide a score.

Our goal is to find the combination rule that puts the techniques in the same, or the closest possible order as the MOS. However, in this initial small-scale experiment we could not ask our interpreters to rank more than a few tens of images, beyond which point their fatigue would rapidly grow and their reliability sharply drop. With this small amount of data, it makes no sense to synthesize the optimal combination rule based on observations since the risk of over-fitting would be extremely high. Therefore, we will set for the less ambitious goal of observing the correlation between the benchmark indicators and the MOS and comparing a few simple combination rules.

4 Experimental Results and Comments

We ran our despeckling algorithm with 20 different parameter sets, filtering both some real-world TerraSAR-X images (©Infoterra GmbH) taken over Rosenheim in Germany, to be analyzed by the interpreters, and the canonical images used to compute the indicators. After a first screening, only ten of the twenty sets of images were retained, because some groups of images were considered too similar to one another to be meaningfully ranked, an interesting fact by itself, considering that the image quality as measured by the benchmark was not at all the same.

In Table 2 we report the ranking obtained for the various indicators. Column 3, for example, shows that set #10 provides the best result in terms of Equivalent number of looks, and set #8 the worst. In the next-to-last column we report the average score which is probably the simplest combination rule that can weight all indicators, and is indeed reported (without comments) also in the cited Prague remote sensing benchmark. In parallel, and in a similar manner, we collected the rankings of the interpreters, not reported here for brevity, and computed their MOS, shown in the last column of the table.

The results seem somewhat dismaying as the two sets of scores are only weakly correlated. In particular, the correlation between the rankings induced by the average on the benchmark and the MOS is just 0.31, definitely too weak to consider the former a good predictor of the latter. Other simple and widespread rules were tested, such as counting the number of times that a set provides the best result, or one of the best, but without significant improvements, so we do not show or comment them here, also because there are indeed many items that deserve analysis and further investigation.

First of all, we want to point out the relatively large correlation between the MOS ranking and some of the rankings induced by selected single indicators, 0.697 for ENL and DG/H, which measure despeckling power, and even 0.724 for the Edge Smearing.

Table 2. Ranking of the parameter sets under different indicators, and mean opinion score

set	MoI/H	ENL	DG/H	DG/D	C_x	ES	FOM	C_{BG}	Average	MOS
1	4	9	9	3	1	3	8	2	4,9	6,3
2	1	3	3	4	4	6	2	3	3,3	5,6
3	7	7	7	9	9	9	6	9	7,9	8,6
4	10	6	6	10	10	10	10	4	8,3	4,2
5	8	8	8	2	2	8	7	7	6,3	7,1
6	2	4	4	1	3	5	3	10	4,0	4,0
7	9	2	2	7	7	1	5	8	5,1	3,7
8	5	10	10	8	5	4	9	6	7,1	4,6
9	3	5	5	5	6	7	4	1	4,5	6,6
10	6	1	1	6	8	2	1	5	3,8	3,9

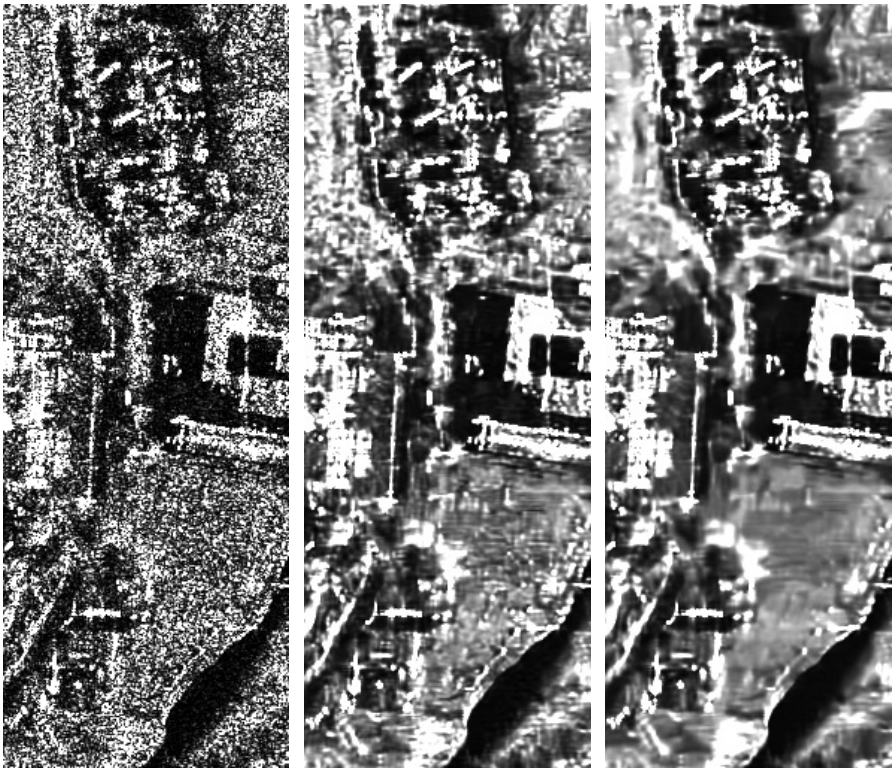


Fig. 3. SAR Image Rosen4, from top to bottom: original, filtered with parameter set #2, set #7

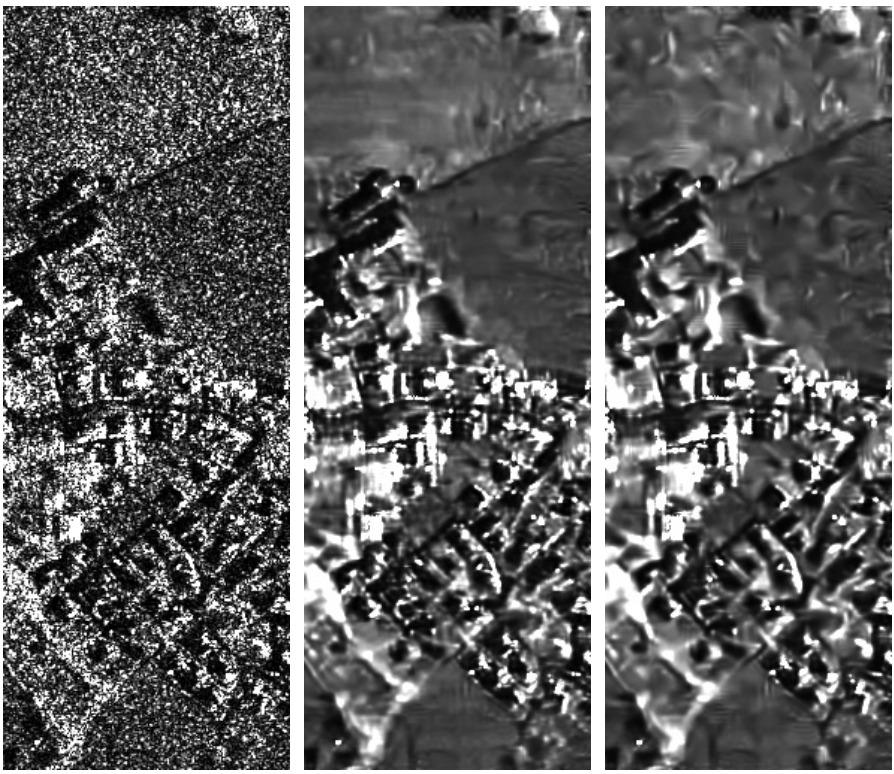


Fig. 4. SAR Image Rosen1, from top to bottom: original, filtered with parameter set #7, set #9

On the contrary, the correlation with the rankings of other indicators are very weak, or even negative. This fact suggests that the interpreters base their decision almost exclusively on a few image features which are more easily spotted. On the other hand, human beings can hardly appreciate the accurate preservation of radiometric accuracy (MoI, C_{BG}) or the preservation of textural features (DG/D, C_x) without some specific analysis tools. Consider the example of Fig.3, portraying a strip of the original Rosen-4 image together with two filtered versions. The benchmark considers set #2 as the best in terms of average performance but the interpreters gave a better MOS to set #7, very likely because it guarantees a better speckle suppression, and this appears clearly in the figure, especially in the vegetated areas. Maybe some fine texture was suppressed by filtering with set #7, but it is really difficult to decide about it without a clean reference.

A further due observation concerns something the interpreters do see, image artifacts, which are instead quite difficult to measure objectively. Some images which do pretty well in terms of benchmark indicators, like those produced with set #9, are affected by despeckling artifacts that impair their quality and are clearly catched by the interpreters but not by benchmark indicators. A good example is shown in Fig.4, with reference to the Rosen-1 image, where the image filtered with set #9 shows many ghost structures in flat areas, almost absent when set #7 is used.

It seems fair to say, in conclusion, that we came far short of our initial goal, finding a predictor of perceived SAR image quality based on benchmark indicators. However, even the limited-scope experiment carried out in this research points out a number of relevant issues that certainly deserve further investigation, from the definition of reliable indicators of the presence of artifacts, to the choice of a better combination rule to keep into due account all aspects of quality. Another obvious issue concerns the experimental setting, since the interpreters should be given tools that enable them to focus on particular details of interest. All this requires a larger-scale study, which we are now designing, with significant resources in terms of personnel and facilities, to consider different despeckling techniques (not just different parameters) and different types of SAR images.

References

1. Vandewalle, P., Kovacevic, J., Vetterli, M.: Reproducible research in signal processing. IEEE Signal Processing Magazine 26, 37–47 (2009)
2. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. on Image Process. 13, 600–612 (2004)
3. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: 8th IEEE International Conference on Computer Vision, ICCV (2001)
4. Scarpa, G., Haindl, M.: Unsupervised texture segmentation by spectral-spatial-independent clustering. In: 18th International Conference on Pattern Recognition, vol. 2, pp. 151–154 (2006)
5. Mikes, S., Haindl, M., Scarpa, G.: Remote sensing segmentation benchmark. In: 7th IAPR International Workshop on Pattern Recognition in Remote Sensing (PRRS 2012), Tsukuba Science City, Japan (November 2012)
6. Handbook on constructing composite indicators. Methodology and user guide, OECD/EC JRC (2008)
7. Van Leeuwen, T.N., Visser, M.S., Moed, H.F., Nederhof, T.J., Van Raan, A.F.J.: The Holy Grail of science policy: exploring and combining bibliometric tools in search of scientific excellence. Scientometrics, 257–280 (2003)
8. Cagnazzo, M., Parrilli, S., Poggi, G., Verdoliva, L.: Cost and advantages of shape adaptive wavelet transform in object-based image coding. EURASIP Journal of Image and Video Processing, 1–13 (2007)
9. Di Martino, G., Poderico, M., Poggi, G., Riccio, D., Verdoliva, L.: Benchmarking framework for SAR despeckling. IEEE Trans. Geosci. Remote Sens. (in Press, 2013)
10. Franceschetti, G., Migliaccio, M., Riccio, D., Schirinzi, G.: SARAS: a SAR raw signal simulator. IEEE Trans. Geosci. Remote Sens. 30(1), 110–123 (1992)
11. Lopes, A., Touzi, R., Nezry, E.: Adaptive speckle filters and scene heterogeneity. IEEE Trans. Geosci. Remote Sens., 992–1000 (1990)
12. Deledalle, C.A., Denis, L., Tupin, F.: Iterative weighted maximum likelihood denoising with probabilistic patch-based weights. IEEE Trans. Image Process. 18, 2661–2672 (2009)
13. Parrilli, S., Poderico, M., Angelino, C.V., Scarpa, G., Verdoliva, L.: A nonlocal approach for SAR image denoising. In: Proc. IGARSS, pp. 726–729 (July 2010)
14. Parrilli, S., Poderico, M., Angelino, C.V., Verdoliva, L.: A nonlocal SAR image denoising algorithm based on LLMMSE wavelet shrinkage. IEEE Trans. Geosci. Remote Sens. 50, 606–616 (2012)
15. Cozzolino, D., Parrilli, S., Scarpa, G., Poggi, G., Verdoliva, L.: Fast adaptive nonlocal SAR despeckling. IEEE Geosci. Remote Sens. Lett. (in Press, 2013)