Multifocus Fusion with Multisize Windows

R. Redondo[a], F. Sroubek[ab], S. Fischer[a] and G. Cristóbal[a]

[a]Instituto de Óptica, CSIC, Serrano 121, 28006 Madrid, Spain;
[b]Academy of Sciences, Pod vodárenskou věží 4, Prague, Czech Republic.

ABSTRACT

The term *fusion* means in general an approach to combine the important information simultaneously from several sources (channels). When we approach image fusion, multiscale transforms (MST) are commonly used as the analyzing tool. It transforms the sources into a space-frequency domain which can be understood as a measure of the saliency (activity level). The criterion to fuse consists of taking the decision to preserve the most salient data from the sources. In order to reduce sensitivity against noise the saliency is often averaged over certain neighborhood (window). However averaging produces that decisions become more fuzzy. Traditionally the size of the neighborhood is chosen fixed according to the level of noise present in the sources, which has to be estimated in advance. This paper proposes a novel technique which combines a set of decreasing averaging windows in order to exploit the advantages of each one. We call it *multisize windows*-based fusion. This technique apart from improving fusion results avoids selecting the neighboring size in advance (and therefore to estimate the level of noise) since it only needs a simple set of windows defined according to image size. We compared it with another technique developed by us called *oriented windows* which, although it consider a fixed neighborhood, adapts the averaging shape to the spatial orientation of the saliency. The specific case of multifocus image fusion is considered for the experiments. The multisize windows technique delivers the best percentage of correct decisions compared with any single fixed window in all the experiments carried out, adding different noise sources (Gaussian, speckle and salt&pepper) with different levels. Although it does not performs better than the oriented window scheme one has to bear in mind that oriented windows are tuned in each case to the best size.

Keywords: multifocus image fusion, wavelet transform

1. INTRODUCTION

The term *fusion* means in general an approach to combine information simultaneously from several sources. An illustration is given by the human system which calls upon its different senses, its memory and its reasoning capabilities to perform deductions from the information it perceives. The goal of image fusion is to integrate complementary multisensor, multitemporal and/or multiview data into a new image containing information, the quality of which cannot be achieved otherwise. The term "quality" depends on the application requirements. The individual images entering the fusion process are called *channels*. Image fusion has been used in many application areas, e.g., in remote sensing and astronomy, in machine vision and mobile robot navigation, in automatic change detection and monitoring of dynamic processes, and last but not least in optical microscopy (multifocus fusion) and medical imaging (multimodal fusion).

Image fusion usually starts with dividing the channels into subregions, calculating a measure of information level in the regions (in the literature often referred to as a *activity level*) (AL) and then utilizing some fusion rules to combine the channels. The channel comparison can be done at different levels of abstraction.[1] The lowest possible is the pixel level, which refers to the merging of measured physical parameters (intensity values of pixels). One step higher is feature-level fusion, which operates on characteristics such as size, shape, edge, contrast and texture. The highest level of abstraction, called decision level fusion, deals with symbolic representations of images. When we talk about image fusion we usually refer to fusion that lies between the pixel and feature level. A common apparatus in image fusion is a multiscale transform (MST), such as the Laplacian pyramid, contrast pyramid, gradient pyramid and wavelet decomposition. Coefficients of MST can be regarded as simple features.

Further author information: (send correspondence to G. Cristóbal)
G. Cristóbal: E-mail: gabriel@optica.csic.es
F. Sroubek: E-mail: filip@optica.csic.es

The measure of information level in the subregion is the critical point in the whole process and several different methods were suggested in the literature. In most of the cases, the AL is proportional to the energy of high frequencies in the channel. It corresponds with an intuitive expectation that high frequencies contain details that are important for our visual perception and understanding of the fused image. Image variance, norm of image gradient, norm of image Laplacian,[2] energy of a Fourier spectrum,[3] image moments,[4] and energy of high-pass bands of a wavelet transform[5–7] belong to the most popular measures of AL.

Another important issue is what technique to employ for dividing the channels into subregions. The simplest but the most common strategy is to use square neighborhoods around each image position. More advanced approaches propose to perform first segmentation of the channels and then use the obtained segments as subregions. At each subregion (or pixel neighborhood), AL's of all channels are compared and the information (pixel values or MST coefficients) of the channel with the highest activity is preserved (*maximum selection rule*). By this process we create a decision map (DM). Alternatively, the first couple of channels with the highest activity can be preserved and their information is averaged. A consistency verification stage follows to prevent occurrence of outlying decisions. One can regard this step as smoothening of the DM. Once the DM is ready, we create the multiscale representation of the fused image and perform the inverse MST. An excellent overview of multiscale image fusion is given in Ref..[8]

The DM plays a crucial role in the whole process since it tells us which information to take at what place. Accurateness of DM is important for valid image reconstruction. One way to increase the accuracy of DM is to integrate into the calculation of AL some additional information about the characteristics of the images in question. We have proposed in Ref.[9] to use oriented neighborhoods that are elongated in the direction parallel to the edge in order to minimize the probability that the neighborhood will cross into another region.

In the sequel we adopt the strategy of simple square neighborhoods (windows) on which AL is calculated. The size of the window depends on the scale of details and on the level of noise in the channels. If the level of noise increases, a larger window is necessary to provide robust AL calculation. Consequently, the DM becomes more fuzzy. Regions close to the decision changes suffer the most from large windows, since here the window intermixes information from regions of potentially different decisions.

The aim of this paper is a methodology that leads to more accurate DM's and that is parameter-free; no tuning parameter that depends on noise (e.g. size of the pixel neighborhood) is necessary. We thus propose to calculate AL's for different window sizes in parallel and take the decision which has the highest level of confidence. We call this technique multisize windows-based fusion. The methodology is general and can be applied to any fusion technique described above. However, in this paper we consider wavelet-based fusion and perform experiments on multifocus data, i.e., we fuse images that depict the same scene but each image was acquired with a different focus length. In this case, AL is often referred to *focus measure* and DM identifies regions in focus. One must assume that there exists a partitioning of the scene into regions and each region is acquired undistorted (in focus) in at least one channel. The identification of undistorted subregions determines the distance of the subregions from camera's (or microscope's) objective lens. Then the distance can be used for surface reconstruction of the measured object. In the case of multifocus fusion, an accurate DM is not only important for valid reconstruction of the fused image, but it is also critical for the surface reconstruction. Erroneous decisions can produce unrealistic peaks and valleys on the surface.

The rest of the paper is organized as follows. The next section reviews the image fusion techniques based on MST. Section 3 introduces the concept of multisize windows and incorporates it in the wavelet-based fusion procedure. In Section 4 experiments on real data under different levels of noise are presented and comparison with a standard wavelet-based fusion is given.

2. MULTISCALE-BASED FUSION

First we give a brief description of the fusion methodology based on the multiscale decompositions. More or less we follow the notation and terminology given in Ref.[6] Let $\mathbf{I}_j$ denote the j-th input channel and $\mathbf{Z}$ denote the fused image. The coefficients of MST can be addressed with a multi-index $\vec{\mathbf{p}} = (m, n, k, l)$, where m, n indicate the spatial position in a given frequency band k and l the decomposition level. In the case of the standard wavelet transform $k = 1, 2, 3$ except the last level, where we have only one low-pass band $W_j(m, n, 1, l_{\max})$

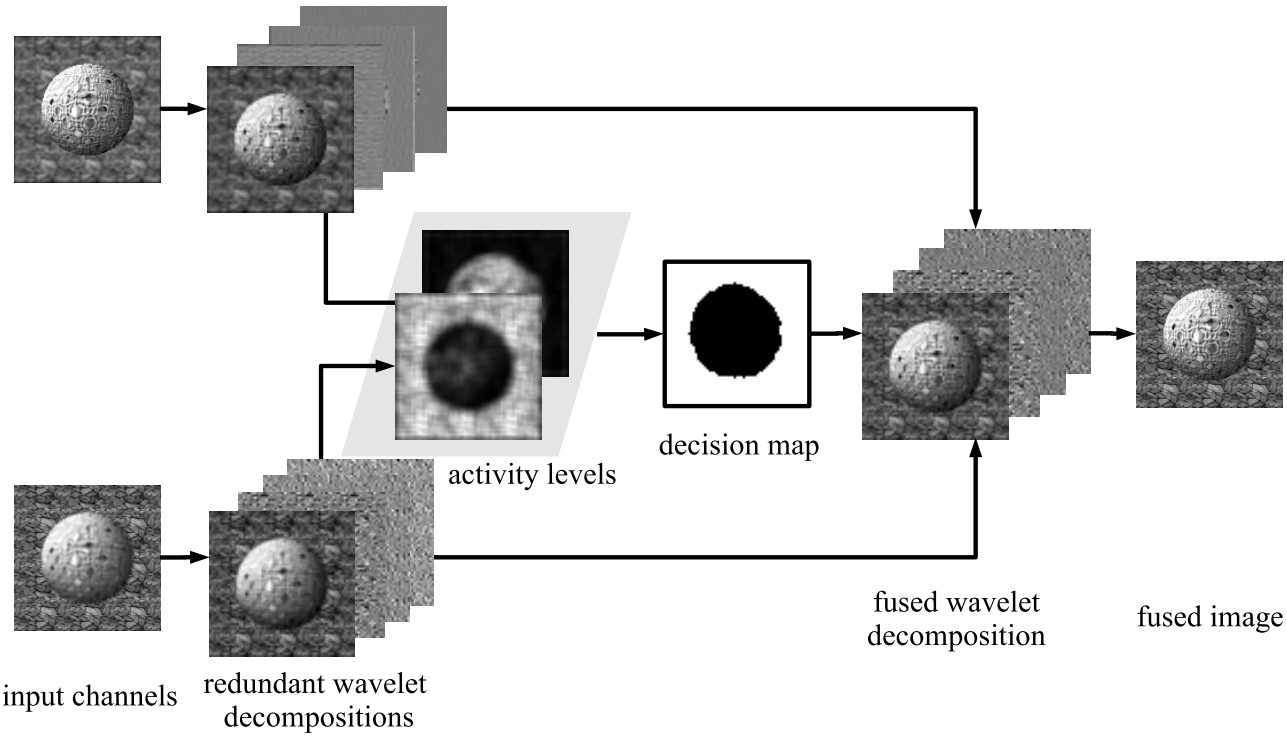

Figure 1. MST fusion steps: Acquire input channel $\mathbf{I}_j$ with different focus settings; perform multiscale decomposition W_j; calculate activity levels A_j; determine the decision map D using the maximum rule; combine the multiscale decompositions using the decision map and create a fused multiscale decomposition $W_{\mathbf{Z}}$; perform inverse multiscale transform to obtain a fused image $\mathbf{Z}$.

(approximation of the signal). We denote the MST transform of $\mathbf{I}_j$ as $W_j(\vec{\mathbf{p}})$. The activity level and the decision map have the same structure as W_j and we denote them as $A_j(\vec{\mathbf{p}})$ and $D(\vec{\mathbf{p}})$ respectively. The complete procedure of multiscale-based fusion is depicted in Fig. 1. In this figure fusion of two input channels with one-level wavelet decomposition is assumed. However, the same scheme applies for any MST and any number of input channels.

The AL of an MST coefficient reflects the local energy in the area spanned by this coefficient in the original image. The AL is related to the absolute or squared value of the corresponding coefficients in the MST domain. The simplest form is to consider each coefficient separately. This is however not robust against noise and therefore window-based AL's were introduced that employ a small (typically 3x3 or 5x5) window centered at the current coefficient position. This approach can be further generalized and leads to weighted averages:

$$A_j^N(\vec{\mathbf{p}}) = \sum_{s,t} w^N(s,t)|W_j(m+s,n+t,k,l)|\,, \tag{1}$$

where w^N is the weighting window of size N that satisfies $\sum_{s,t=-N/2}^{N/2} w(s,t) = 1$.

Instead of averaging one can use rank filters. The popular choice here is to pick the maximum absolute value on the given neighborhood as our AL. Another option is to segment the input channels and calculate one AL for each image segment. In Eq.1 the weight w then corresponds to the characteristic function of the segment. This approach is referred to region-based activity measurement.

To build DM for the given window size N, the most common scheme is to apply the maximum rule to AL's. Formally, we can write

$$D^N(\vec{\mathbf{p}}) = \arg\max_j \big(A_j^N(\vec{\mathbf{p}})\big)\,. \tag{2}$$

The decision map D has the same structure as W or A. It contains indices of the input channels and determines which MST coefficients to use at what place. In addition, if the focus setting for each channel j is known, D

corresponds to the depth map and it can be used for the surface reconstruction. Using the decision map, the composite MST representation $W_{\mathbf{Z}}$ of the fused image $\mathbf{Z}$ is given by

$$W_{\mathbf{Z}}(\vec{\mathbf{p}}) = W_{D(\vec{\mathbf{p}})}(\vec{\mathbf{p}})\,. \tag{3}$$

The maximum rule considers at each position $\vec{\mathbf{p}}$ only the strongest MST coefficient and thus only one channel. Another possibility is to perform weighted averaging of the MST coefficients using weights proportional to AL's. However in the case of multifocus fusion, this combination scheme lacks any scientific support. Since we assume that each position (pixel) in the original image is acquired undistorted in at least one channel, only the maximum rule sounds perfectly plausible. The same holds true for any coefficient grouping method. One should avoid different decisions at different levels l and frequency bands k of MST. Therefore, we implement only one-level redundant wavelet decomposition (one low-pass and three high-pass bands, each of the same size as the input channel), calculate A as a maximum of AL's of three high-pass bands and use this A also for the low-pass band, i.e.,

$$\dot{A}_j^N(m,n) = \max_k \big[\sum_{s,t} w^N(s,t)|W_j(m+s,n+t,k,1)|\big]\,. \tag{4}$$

Note that for a given N we have for each input channel j exactly one AL of the same size as the channel and one DM.

3. MULTISIZE WINDOWS-BASED FUSION

As previously mentioned the area $N \times N$ of the window which contributes to the frequency analysis in a certain position (pixel) constitutes a critical factor in terms of noise robustness and space location. But both features exclude each other. Large areas are often used to reduce noise impact, however it implies higher delocalization because outer signals in the window could be totally different to middle ones, what could mislead the analysis. On the other hand, a tiny averaging window narrows delocalization, but increases noise vulnerability. Heisenberg's Uncertainty Principle is often referred in order to explain such a phenomenon. The robustness against noise is important all over the input image, but the importance of localization becomes specially critical where a transition between two regions of different decision (in multifocus imaging two different focus planes) because different frequency responses are being mixed up. To handle this uncertainty about either small or large windows, information provided by several window sizes therefore needs to be properly applied in a space-variant window combination.

Given a certain set of window sizes Ω, the aim is to trust in the largest averaging window wherever its location was confident, otherwise a smaller window must be applied. Inside the regions with the same decision fused pixels come from the same channel (the DM is uniform) and close the regions borders pixels of different channels lie side by side. Therefore it seems logical that such a confidence can be thought as the highest number of neighboring pixels that come from the same channel. By normalizing the number relative to window area we obtain a value $p_j^N(m,n)$ from 0 (uncertain) to 1 (certain) which tell us the spatial confidence of each window on each pixel. This value can be seen as a density of probability in a certain neighborhood. Now for each pixel the input channel more probable is considered and those which are over a certain confidence threshold θ (1- strict 0-relaxed) merge the DM following its neighbor majority and the remaining pixels will be judged in a next step by applying a smaller window.

Once defined the criteria (to combine wavelet domain by using hierarchically different window sizes according to a certain confidence value), it remains to formulate the procedure.

1. Let channel index be $j = 1..J$.
2. Create empty decision map $\hat{D}(m,n) = 0$.
3. $N = \max(\Omega)$.
4. For every $\hat{D}(m,n) = 0$ do
   if $\max_j \big[p_j^N(m,n)\big] > \theta_N$ then
   $\hat{D}(m,n) = \arg\max_j \big[p_j^N(m,n)\big]$;

5. Take the next largest $N \in \Omega$ and repeat 4 (until the smallest $N \in \Omega$ was selected).

Note that if certain position of $\hat{D}(m,n)$ has been already selected then it is not further computed anymore. The thresholds θ_N (now in plural because depends on the window area) governs in some sense the amount of decisions preserved from each D^N. A graphical way of understanding how this procedure works in reality could be a "filling-in" process, first by considering decisions from the largest windows all over the inner part on focus regions and the more it approaches to a border between two on focus regions from different channels (a focal plane transition) the more decisions from the smallest window begin to be considered because the largest ones are rejected more and more unable to reach the confidence.

However, depending on the confidence thresholds not all the pixels can be merged at the end of such a hierarchical process, that is, some pixels are below θ_N for every N. Indeed the higher are θ_N the fewer decisions are made. In general these undecided pixels are just spread along boundaries delimiting focal changes, where the uncertainly is maximum. The strategy to follow now (in such remains) consists of performing a linear combination over every AL in order to decrease uncertainly:

$$\tilde{D}(m,n) = \arg\max_j \big[\sum_N (\dot{A}_j^N(m,n)\big]. \tag{5}$$

Therefore the final DM remains as follows:

$$D(m,n) = \hat{D}(m,n) + \tilde{D}(m,n). \tag{6}$$

The size of the averaging window has been traditionally chosen with regard to the level of noise in input channels $\mathbf{I}_j$, which means that the size has to be fixed in advance, in other words, it has to be predicted. One advantage of this multisize windows scheme, besides refining the appearance of DM as we will see, is its independence on the window size. It is true that a set of sizes Ω and a set of thresholds θ_N should be previously defined but such sets depend almost exclusively on the size of the input channels (which can be straightforward automatized).

4. ASSESSMENT

The data set consists of images acquired with a standard digital camera in a laboratory environment (see Fig.2(a)-(b)). Apart from blurred versions we are able to acquire an image which is sufficiently sharp everywhere to approximate a "ground truth" image (Fig.2(c)) and estimate an ideal decision map (Fig.2(d)). We can then calculate the percentage of correct decisions (PCD) to evaluate the quality of DM. The evaluation measures are defined as follows:

$$\text{PCD} = 100\frac{N_c}{N_t}\,, \tag{7}$$

where N_c is the number of correct decisions in the calculated DM and N_t is the total number of decisions, i.e., the size of the image.

Hereafter the set of windows is $\Omega = [15, 11, 9, 7]$ and confidence thresholds $\theta_N = [0.8, 0.8, 0.7, 0.6]$. This set of values were adjusted relative to the size of input channels $\mathbf{I}_j$ and validated experimentally. Subsequently they were not changed anymore. However, one requirement verified experimentally is that windows should not be neither too large nor too small, which on the other hand is reasonable. The second one is that confidence thresholds of the smallest windows should be more relaxed because they perform in regions of focal changes where they are the most reliable, which on the other hand is again reasonable. If these two guidelines are followed the method behaves quite stable.

The Fig.3 depicts a graphical example of how the size of the averaging window affects the DM appearance. Cases (a)-(b) correspond to decisions taken by windows of size 7 and 15, respectively. The decisions in (a), the smallest size, are more random, however it is more precise around the transition between the Indian and the background. Exactly the opposite happens in (b), there is less randomness in decisions but boundaries between the Indian and the background are now more fuzzy. Such examples show up visually the advantages and disadvantages of each window and how their efficiency fluctuates with respect to the location. The DM in (c) is the result of combining the set of four windows Ω with the confidence thresholds θ_N. Now decisions are

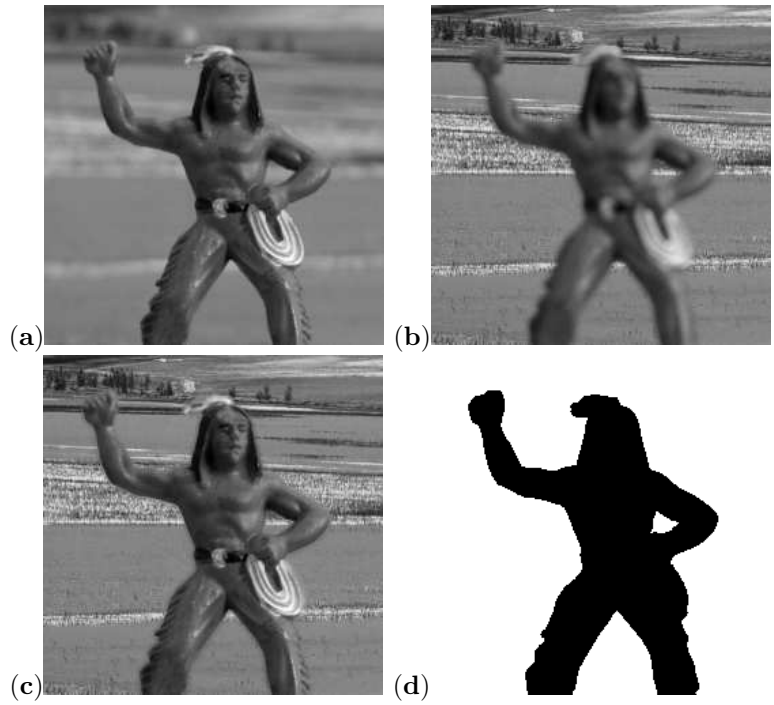

Figure 2. Laboratory experiment: Images 250×250 depict a two-plane scenario. (a)-(b) Two input images with the Indian in focus and a photography as background in focus, respectively. (c) Ideal image that is sharp everywhere was acquired with a large aperture. (d) Mask of the Indian that defines the ideal decision map DM.

more robust and the Indian contour is more precise at the same time. Even better results in terms of PCD were achieved for oriented windows.[9] However, it is important to note that, contrary to the proposed method, this technique has a tuning parameter (variance of oriented windows) that is noise dependent and that the best possible value of the parameter was used in this case. In real applications it is difficult to estimate this parameter and therefore parameter-free techniques, such as the proposed one, prove their superiority. Fig.4 shows more explicitly the discrepancies between decision maps in Fig.3 and the ideal decision map in Fig.2(d).

A simple visual inspection already shows signs of improvement of the proposed method. However, in order to assess thoroughly the performance of the multisize windows-based fusion. The data set is exposed under three types of noise of different strength. The three types of noise evaluated are Gaussian, speckle and salt&pepper, whose formulation is not specified here but can be found in the literature. All of them are generated in a pixels-independent manner so that they are white noise. The first two appear commonly in many natural process, the last one is usually recreated under laboratory conditions. The set of variances is the appropriate to test the method along a wide range of conditions, from weak to severe. Decision maps are totally wrong beyond variances over 10^{-2} so there is no sense to be considered and they do not improve below 10^{-6}. Moreover, 100 instantiations are taken from each noise evaluation in order to avoid randomness.

Fig.5 shows the result in terms of PCD. The smallest window $N = 7$ is the most vulnerable in presence of noise and its performance decays drastically when noise increases, in spite of performing the best when without noise. There again $N = 15$ decays less but without presence of noise its poor localization becomes important. All the lines crosses somewhere near to such a drastic drop what means once again the appropriate window size changes according to the noise and space location in the input images.

The graphs also corroborate the superiority of the multisize windows. By utilizing different sizes in an appropriate space order we take double advantage, precision and noise robustness, which is demonstrated along all experiments carried out. The multisize windows-based fusion achieves PCD higher than any other single

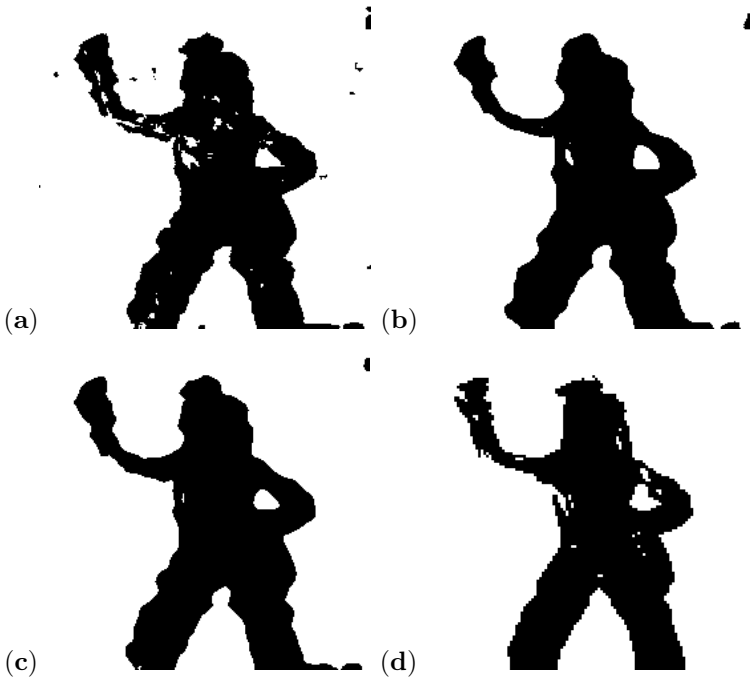

Figure 3. Decision maps DM examples. Cases (a)-(b) correspond to DM using windows of size 7 and 15, respectively. Case (c) corresponds to DM by using multisize windows combination with $\Omega = [15, 11, 9, 7]$ and $\theta_N = [0.8, 0.8, 0.7, 0.6]$, and (d) corresponds to DM by using multioriented window approach.

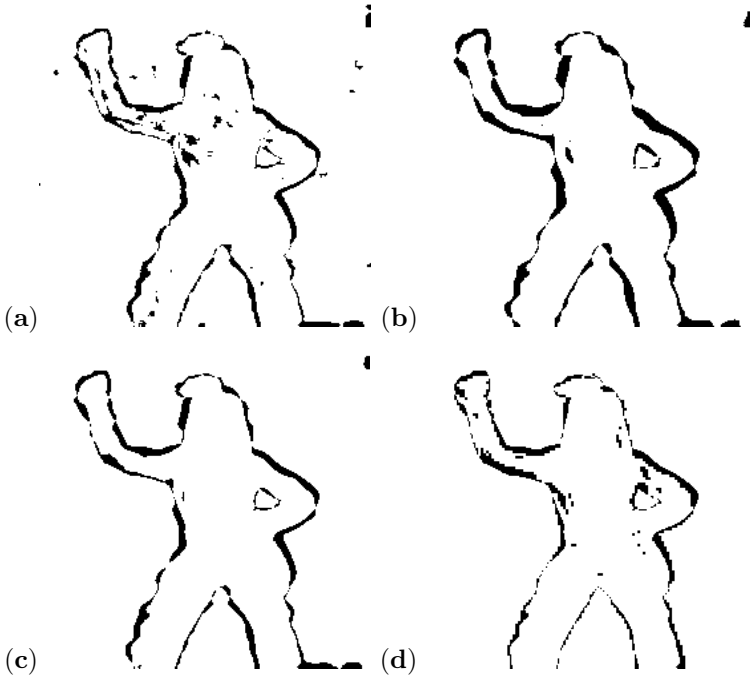

Figure 4. Discrepancies between DM in Fig.3 and the ideal decision map in Fig.2(d). Labels (a)-(d) have exactly the same correspondence.

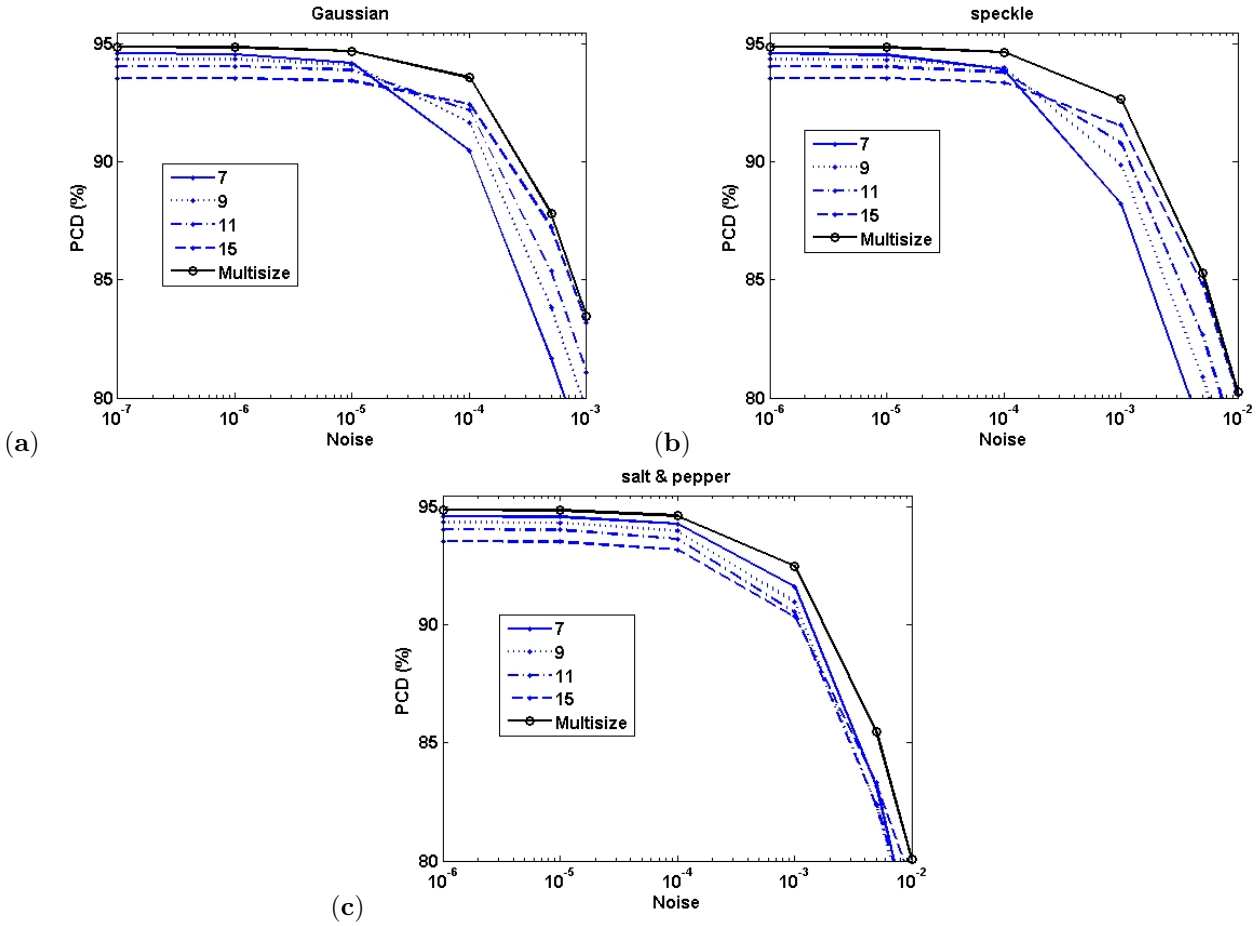

Figure 5. Percentage of Correct Decision (PCD) under (a) Gaussian, (b) speckle and (c) salt&pepper noises. Horizontal axis reflects the power of noise (its variance). Windows tested are $\Omega = [15, 11, 9, 7]$ and confidence thresholds are $\theta_N = [0.8, 0.8, 0.7, 0.6]$.

window and that difference holds almost constantly in a wide range of levels of noise for the three types of noises tested here.

A common behavior of multisize and the single size averaging window is the drastic drop of PCD when the level of Gaussian noise goes beyond 10^{-4} and 10^{-3} for speckle and salt&pepper, which corresponds when noise becomes more visible. As a result we can conclude that Gaussian is the most aggressive and speckle noise and salt&pepper delivers quite similar rates (10dB over Gaussian) although speckle is rather less aggressive.

5. CONCLUSIONS

Traditionally multiscale image fusion has employed a window of fixed size to average the noise impact. However we have showed how the analysis may change significantly depending on its area. In order to mitigate this problem we propose a multisize windows-based fusion by means of a hierarchical algorithm which combines windows from large to small ones. The multisize scheme avoids large windows intermixes information from regions of potentially different decisions and exploit its better noise robustness on regions where decisions are uniform (come from the same channel).

For the experiments the specific case of multifocus image fusion has been considered. The experiments carried out show the proposed technique performs better than the best single window size and it is more robust against

noise, at least for the three types of noise tested (Gaussian, speckle and salt&pepper). Apart from improving the fusion results we want to stress there is no tuning parameter (window size) that depends on noise.

We also compared it with oriented window method which adapts the averaging shape to the spatial orientation of the saliency. Although it performs better than the multisize windows scheme one has to bear in mind that oriented window technique is tuned in each case to the best window size. However, they do not exclude each other so that they can be further combined, which will the subject of our future research.

In multifocus fusion we used only one-level wavelet decomposition with the multisize windows. In other fusion applications, such as multimodal imaging, more levels can improve results. Further work will thus consider a different strategy by applying the multisize window method to all the levels of a multiscale pyramid.

ACKNOWLEDGMENTS

This work has been supported in part by the grants TEC 2004-00834, the G03/185-IM3 medical imaging thematic network and the bilateral project: 2004CZ0009 CSIC-Academy of Sciences of the Czech Republic, and No. 102/04/0155, No. 202/05/0242 of the Grant Agency of the Czech Republic. RR and SF are supported by CSIC-I3P and MEC-FPU fellowships, respectively. F. Sroubek is supported by the Spanish States Secretary of Education and Universities fellowship.

REFERENCES

1. H. Wang, "A new multiwavelet-based approach to image fusion," *Journal of Math. Imaging and Vision* **21**, pp. 177–192, 2004.
2. M. Subbarao, T. Choi, and A. Nikzad, "Focusing techniques," *Optical Eng.* **32**, pp. 2824–2836, 1993.
3. M. Subbarao and J. K. Tyan, "Selecting the optimal focus measure for autofocusing and depth-from-focus," *IEEE Trans. Pattern Analysis and Machine Intelligence* **20**, pp. 864–870, 1998.
4. Y. Zhang, Y. Zhang, and C. Wen, "A new focus measure method using moments," *Image and Vision Computing* **18**, pp. 959–965, 2000.
5. H. Li, B. Manjunath, and S. Mitra, "Multisensor image fusion using the wavelet transform," *Graphical Model and Image Processing* **57**, pp. 235–245, May 1995.
6. Z. Zhang and R. Blum, "A categorization of multiscale-decomposition-based image fusion schemes with a performance study for a digital camera application," in *Proceedings of the IEEE*, **87**, pp. 1315–1326, Aug. 1999.
7. J. Kautsky, J. Flusser, B. Zitová, and S. Šimberová, "A new wavelet-based measure of image focus," *Pattern Recognition Letters* **23**, pp. 1785–1794, 2002.
8. G. Piella, "A general framework for multiresolution image fusion: from pixels to regions," *Information Fusion* **4**, pp. 259–280, 2003.
9. F. Sroubek, S. Gabarda, R. Redondo, S. Fischer, and G. Cristobal, "Multifocus fusion with oriented windows," in *Bioengineered and Bioinspired Systems II, Proceedings*, SPIE, (Bellingham), May 2005.