ELSEVIER

Pattern Recognition Letters 23 (2002) 985–997

Pattern Recognition Letters

www.elsevier.com/locate/patrec

Multifocus image fusion using artificial neural networks

Shutao Li [a,b,*], James T. Kwok [b], Yaonan Wang [a]

[a] *College of Electrical and Information Engineering, Hunan University, Changsha 410082, People's Republic of China*

[b] *Department of Computer Science, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong*

Received 30 March 2001; received in revised form 21 June 2001

Abstract

Optical lenses, particularly those with long focal lengths, suffer from the problem of limited depth of field. Consequently, it is often difficult to obtain good focus for all objects in the picture. One possible solution is to take several pictures with different focus points, and then combine them together to form a single image. This paper describes an application of artificial neural networks to this pixel level multifocus image fusion problem based on the use of image blocks. Experimental results show that the proposed method outperforms the discrete wavelet transform based approach, particularly when there is a movement in the objects or misregistration of the source images. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Image fusion; Neural networks; Probabilistic neural networks; Radial basis function networks

1. Introduction

Optical lenses, particularly those with long focal lengths, suffer from the problem of limited depth of field. Consequently, the image obtained will not be in focus everywhere, i.e., if one object in the scene is in focus, another one will be out of focus. A possible way to alleviate this problem is by image fusion (Zhang and Blum, 1999), in which several pictures with different focus points are combined to form a single image. This fused image will then hopefully contain all relevant objects in focus (Li et al., 1995; Seales and Dutta, 1996).

The simplest image fusion method just takes the pixel-by-pixel average of the source images. This, however, often leads to undesirable side effects such as reduced contrast. In recent years, various alternatives based on multiscale transforms have been proposed. The basic idea is to perform a multiresolution decomposition on each source image, then integrate all these decompositions to produce a composite representation. The fused image is finally reconstructed by performing an inverse multiresolution transform. Examples of this approach include the Laplacian pyramid (Burt and Andelson, 1983), the gradient pyramid (Burt and Kolczynski, 1993), the ratio-of-low-pass pyramid (Toet et al., 1989) and the morphological pyramid (Matsopoulos et al., 1994). More recently, the discrete wavelet transform (DWT) (Chipman et al., 1995; Koren et al., 1995; Li et al.,

[*] Corresponding author.
E-mail address: shutao_li@yahoo.com.cn (S. Li).

PII: S0167-8655(02)00029-6

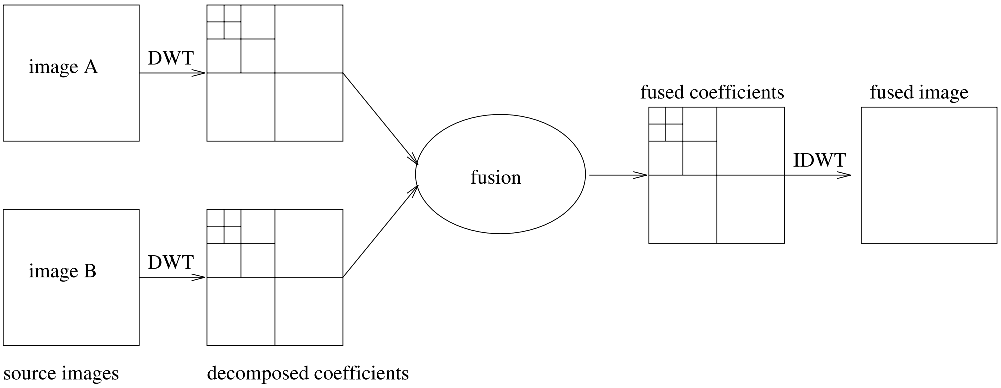

Fig. 1. Image fusion based on the DWT.

1995; Yocky, 1995, 1999) has also been used. In general, DWT is superior to the previous pyramid-based methods (Li et al., 1995). First, the wavelet representation provides directional information while pyramids do not. Second, the wavelet basis functions can be chosen to be orthogonal and so, unlike the pyramid-based methods, DWT does not carry redundant information across different resolutions. Upon fusion of the wavelet coefficients, the maximum selection rule is typically used, as large absolute wavelet coefficients often correspond to salient features in the images. Fig. 1 shows a schematic diagram for the image fusion process based on DWT.

While these methods often perform satisfactorily, their multiresolution decompositions and consequently the fusion results are shift-variant because of an underlying down-sampling process. When there is a slight camera/object movement or when there is misregistration of the source images, their performance will thus quickly deteriorate. One possible remedy is to use the shift-invariant discrete wavelet frame transform (Unser, 1995). However, the implementation is more complicated and the algorithm is also more demanding in terms of both memory and time.

In this paper, we propose a pixel level multifocus image fusion method based on the use of image blocks and artificial neural networks. The implementation is computationally simple and can be realized in real-time. Experimental results show that it outperforms the DWT-based method. The rest of this paper is organized as follows. The proposed fusion scheme will be described in Section 2. Experiments will be presented in Section 3, and the last section gives some concluding remarks.

2. Neural network based multifocus image fusion

Fig. 2 shows a schematic diagram of the proposed multifocus image fusion method. Here, we consider the processing of just two source images, though the algorithm can be extended straightforwardly to handle more than two. Moreover, the source images are assumed to have been registered.

The basic fusion algorithm will be described in Section 2.1. The input features to the neural networks will be discussed in Section 2.2. Section 2.3 contains a brief introduction to the two neural network models that will be used in the experiments.

2.1. The basic algorithm

The algorithm first decomposes the source images into blocks. Given two of these blocks (one from each source image), a neural network is trained to determine which one is clearer. Fusion then proceeds by selecting the clearer block in constructing the final image. As mentioned in Section 1, the fusion result of DWT is shift-dependent. The use of image blocks, on the other

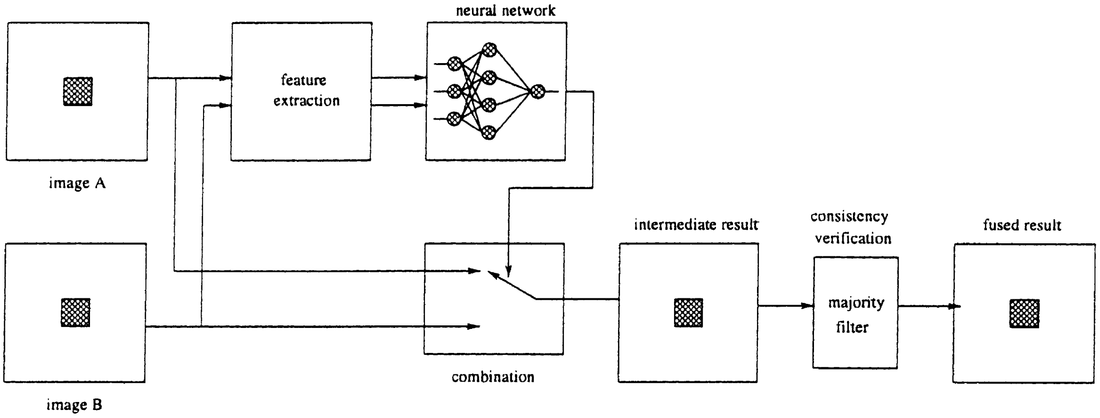

Fig. 2. Schematic diagram of the proposed fusion method.

hand, avoids this problem. Even if there is object movement or misregistration in the source images, each object will still be in better focus in one of the source images. Thus, in the fused image, all the blocks covering a particular object will come from the same source image and hence its clarity will not be affected by any misregistration problem.

In detail, the algorithm consists of the following steps:

1. Decompose the two source images A and B into $M \times N$ blocks. Denote the ith image block pair by A_i and B_i, respectively.
2. From each image block, extract three features that reflect its clarity (details in Section 2.2). Denote the feature vectors for A_i and B_i by $(SF_{A_i}, VI_{A_i}, EG_{A_i})$ and $(SF_{B_i}, VI_{B_i}, EG_{B_i})$, respectively.
3. Train a neural network to determine whether A_i or B_i is clearer. The difference vector $(SF_{A_i} - SF_{B_i}, VI_{A_i} - VI_{B_i}, EG_{A_i} - EG_{B_i})$ is used as input, and the output is labeled according to

$$\text{target}_i = \begin{cases} 1 & \text{if } A_i \text{ is clearer than } B_i, \\ 0 & \text{otherwise.} \end{cases}$$

4. Perform testing of the trained neural network on all image block pairs obtained in Step 1. The ith block, Z_i, of the fused image is then constructed as

$$Z_i = \begin{cases} A_i & \text{if out}_i > 0.5, \\ B_i & \text{otherwise,} \end{cases} \tag{1}$$

where out_i is the neural network output using the ith image block pair as input.
5. Verify the fusion result obtained in Step 4. Specifically, if the neural network decides that a particular block is to come from A but with the majority of its surrounding blocks from B, this block will be switched to come from B. In the implementation, a majority filter with a 3 × 3 window is used.

2.2. Feature extraction

In principle, in Step 3 of the algorithm above, the two $M \times N$ blocks A_i and B_i can be fed directly into a neural network for discrimination. But, in practice, using such a long feature vector (of size

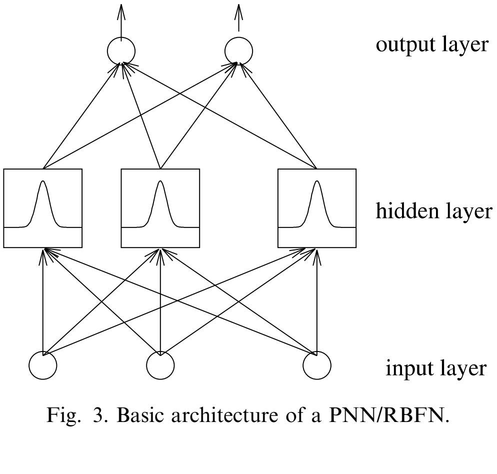

Fig. 3. Basic architecture of a PNN/RBFN.

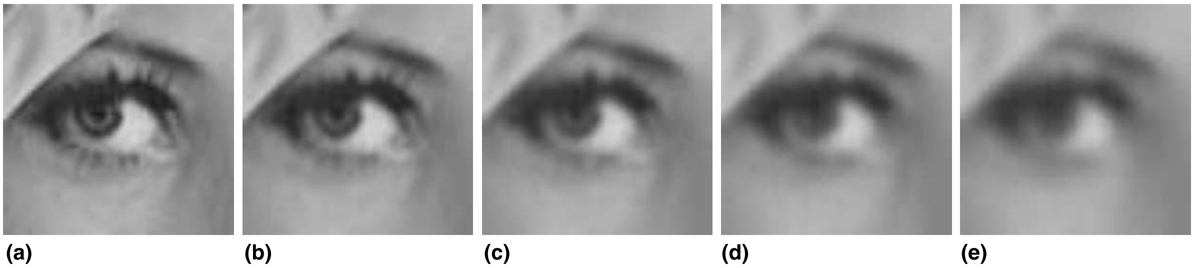

Fig. 4. Original and blurred versions of an image block extracted from "Lena". (a) Original; (b) radius = 0.5; (c) radius = 0.8; (d) radius = 1.0; (e) radius = 1.5.

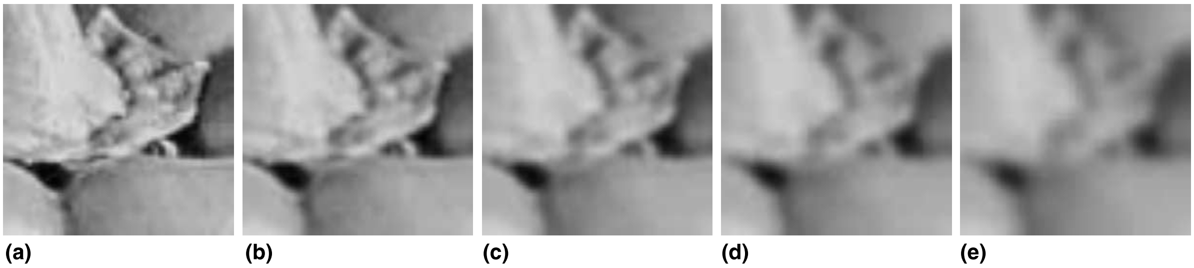

Fig. 5. Original and blurred versions of an image block extracted from "Peppers". (a) Original; (b) radius = 0.5; (c) radius = 0.8; (d) radius = 1.0; (e) radius = 1.5.

2*MN*) can lead to slow convergence and poor generalization performance (*curse of dimensionality* (Bellman, 1961)). Feature extraction is thus usually imperative.

In this paper, we extract three features from each decomposed image block to represent its clarity. These are the spatial frequency (Section 2.2.1), visibility (Section 2.2.2) and an edge feature (Section 2.2.3). The effectiveness of these three features in representing image clarity will be experimentally demonstrated in Section 3.1.

2.2.1. Spatial frequency (SF)

Spatial frequency is used to measure the overall activity level of an image (Eskicioglu and Fisher, 1995). For an $M \times N$ image F, with the gray value at pixel position (m,n) denoted by $F(m,n)$, its spatial frequency is defined as

$$SF = \sqrt{RF^2 + CF^2},$$

where RF and CF are the row frequency

$$RF = \sqrt{\frac{1}{MN}\sum_{m=1}^{M}\sum_{n=2}^{N}(F(m,n) - F(m,n-1))^2},$$

and column frequency

$$CF = \sqrt{\frac{1}{MN}\sum_{n=1}^{N}\sum_{m=2}^{M}(F(m,n) - F(m-1,n))^2},$$

respectively.

Table 1
Feature values for the image blocks in Fig. 4

	Fig. 4(a)	Fig. 4(b)	Fig. 4(c)	Fig. 4(d)	Fig. 4(e)
SF	16.10	12.09	9.67	8.04	6.49
VI	0.0069	0.0066	0.0062	0.0059	0.0055
EG	269	243	225	183	181

Table 2
Feature values for the image blocks in Fig. 5

	Fig. 5(a)	Fig. 5(b)	Fig. 5(c)	Fig. 5(d)	Fig. 5(e)
SF	28.67	17.73	12.98	10.04	7.52
VI	0.0067	0.0063	0.0060	0.0057	0.0054
EG	329	310	274	260	216

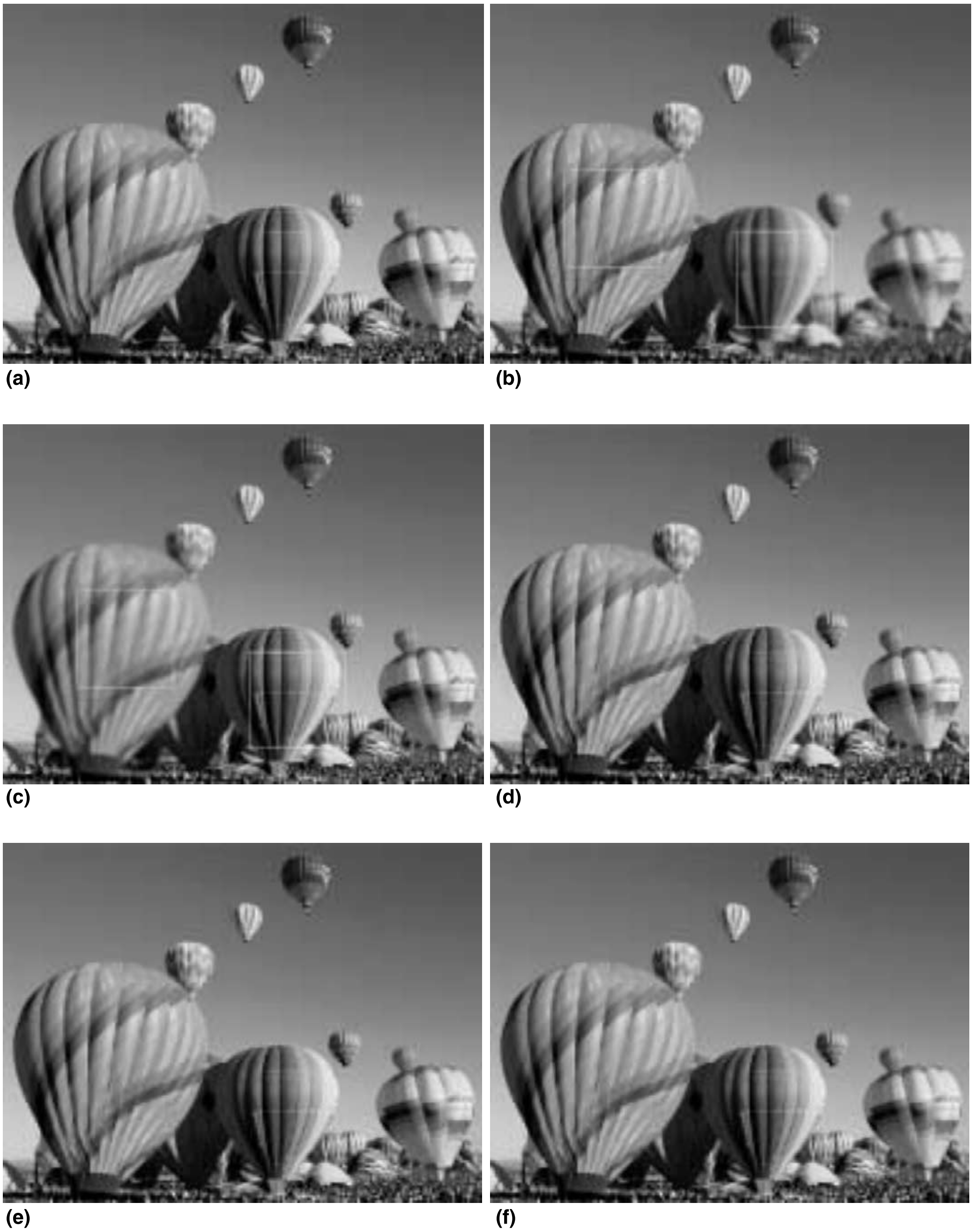

Fig. 6. The “Balloon” reference image, blurred images and fusion results. The training set is selected from regions marked by the rectangles in (b) and (c). (a) Reference image (all in focus); (b) focus on the left; (c) focus on the right; (d) fused image using DWT; (e) fused image using PNN and (f) fused image using RBFN.

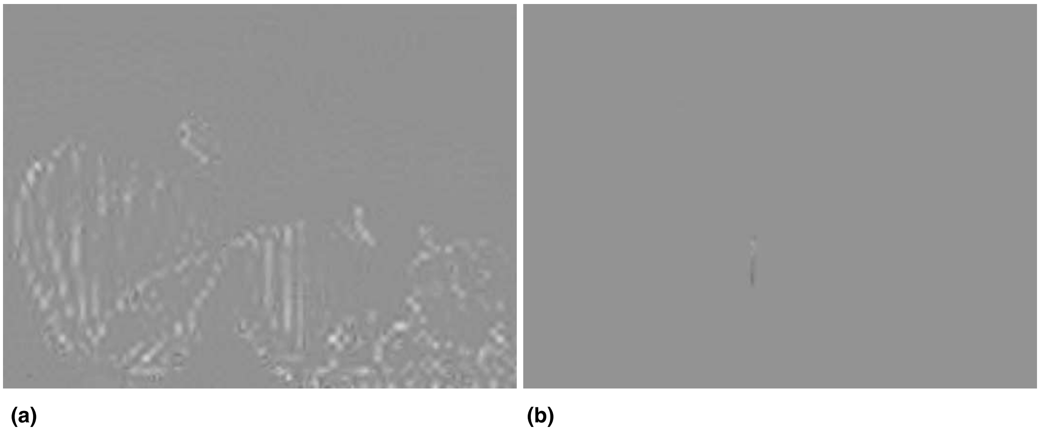

Fig. 7. Differences between the fused images in Fig. 6(d), (e) and source image Fig. 6(a). (a) Difference between the fused image using DWT (Fig. 6(d)) and source image Fig. 6(a). (b) Difference between the fused image using PNN (Fig. 6(e)) and source image Fig. 6(a).

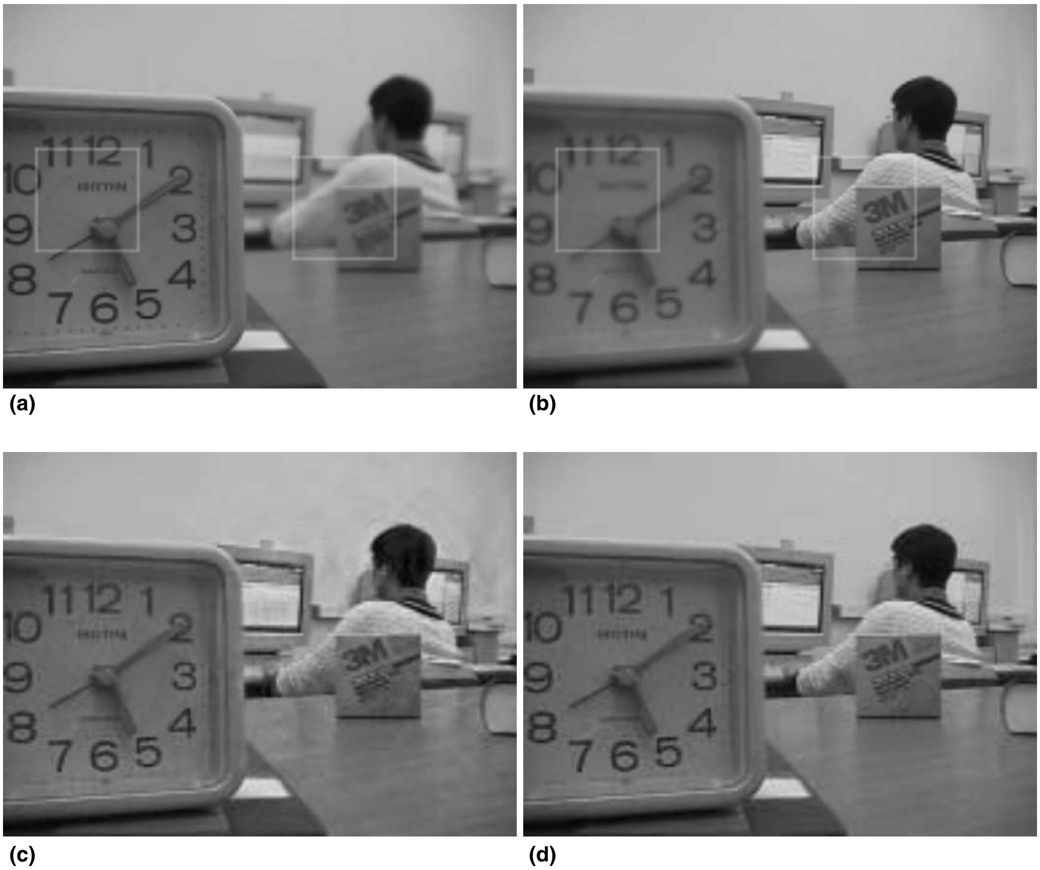

Fig. 8. The "Lab" source images and fusion results. The training set is selected from regions marked by the rectangles in (a) and (b). (a) Focus on the clock; (b) focus on the student; (c) fused image using DWT (db40, level = 5) and (d) fused image using PNN.

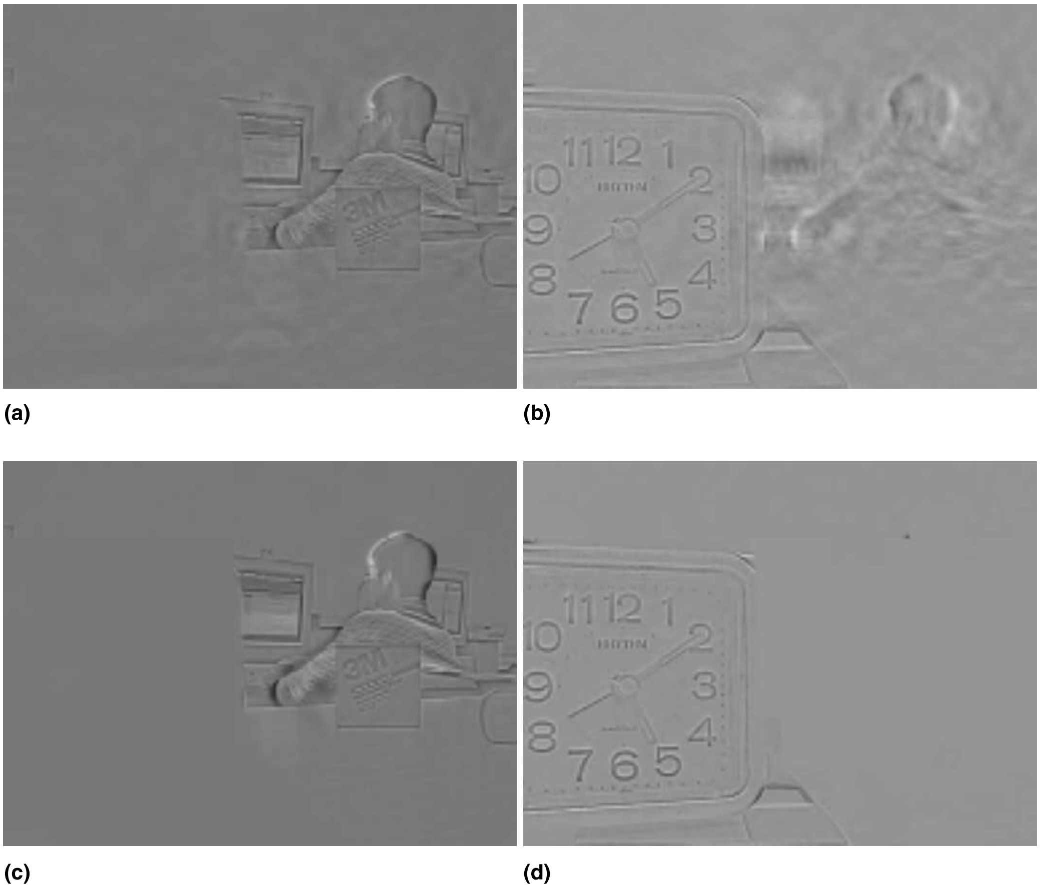

Fig. 9. Differences between the fused images in Fig. 8(c), (d) and source images in Fig. 8(a), (b). (a) Difference between the fused image using DWT (Fig. 8(c)) and source image Fig. 8(a). (b) Difference between the fused image using DWT (Fig. 8(c)) and source image Fig. 8(b). (c) Difference between the fused image using PNN (Fig. 8(d)) and source image Fig. 8(a). (d) Difference between the fused image using PNN (Fig. 8(d)) and source image Fig. 8(b).

2.2.2. Visibility (VI)

This feature is inspired from the human visual system, and is defined as (Huang et al., 1999)

$$VI = \sum_{m=1}^{M} \sum_{n=1}^{N} \frac{|F(m,n) - \mu|}{\mu^{\alpha+1}}, \tag{2}$$

where μ is the mean intensity value of the image, and α is a visual constant ranging from 0.6 to 0.7.

2.2.3. Edge feature (EG)

This feature is based on the number of edges extracted from the image. Intuitively, for images of comparable complexity, a clearer image will have more edges. Here, we first apply the Canny edge detector (Canny, 1986) to each decomposed block. The total number of one's in the resultant binary image block is then taken as the edge feature.

2.3. Artificial neural networks

Many neural network models have been proposed for tackling a diverse range of problems (Bishop, 1995; Hertz et al., 1991), including pattern classification, function approximation and regression. The fusion problem we examine here can be considered as a classification problem. In this paper, we consider two closely related neural network models, namely, the probabilistic neural network (PNN) (Specht, 1990) and radial basis function network (RBFN) (Moody and Darken, 1998).

The idea underlying both networks is to overlap localized receptive fields of the hidden units to create arbitrarily complex nonlinearities. The typical architecture consists of one hidden layer and one output layer (Fig. 3). Each hidden unit corresponds to a basis or kernel function of the input vector $\boldsymbol{x}$, and is usually of the Gaussian form:

$$h(\boldsymbol{x}) = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{c}\|^2}{\sigma^2}\right). \tag{3}$$

Here, $\boldsymbol{c}$ is the position of the hidden unit and σ is a user-defined width that controls its spread. For PNNs, a hidden unit is positioned at every training data point, whereas for RBFNs, the number of hidden units is usually smaller. The network output is simply a linear summation of these basis functions.

3. Experiments

3.1. Demonstration of the effectiveness of the features

In this section, we first experimentally demonstrate the effectiveness of the three features pro-

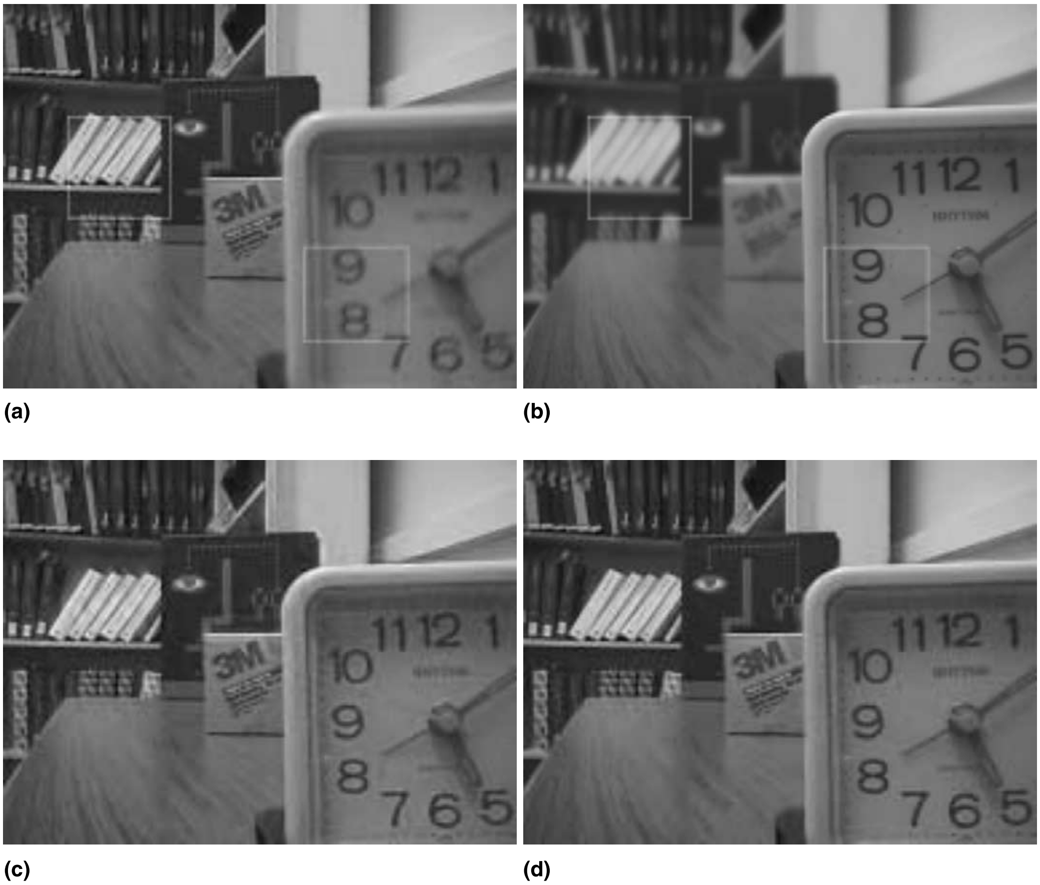

Fig. 10. The “Disk” source images and fusion results. The training set is selected from regions marked by the rectangles in (a) and (b). (a) Focus on the left; (b) focus on the right; (c) fused image using DWT (db8, level = 5) and (d) fused image using PNN.

posed in Section 2.2 (namely SF, VI and EG) in representing the clarity level of an image. An image block of size 64×64 (Fig. 4(a)) is extracted from the “Lena” image. Fig. 4(b)–(e) show the degraded versions by blurring with a Gaussian of radius 0.5, 0.8, 1.0 and 1.5, respectively. As can be seen from Table 1, when the image becomes more blurred, all the three feature values diminish accordingly. Another experiment on an image block extracted from the “Peppers” image produces similar results (Fig. 5 and Table 2). These suggest that all the three features can be used to reflect image clarity.

3.2. Quantitative evaluation of the fusion method

3.2.1. Experimental setup

Experiment is performed on an 256-level image of size 480×640 (Fig. 6(a)), with good focus everywhere. We artificially produce a pair of out-of-focus images, by blurring the left part to obtain the image in Fig. 6(b), and then blurring the right part to produce the image in Fig. 6(c). Blurring is accomplished by using a Gaussian of radius 2.

Image blocks of size 32×32 are used. Two pairs of regions, each containing 30 image block pairs, are selected from the two out-of-focus images. In 15 of these block pairs, the first image is

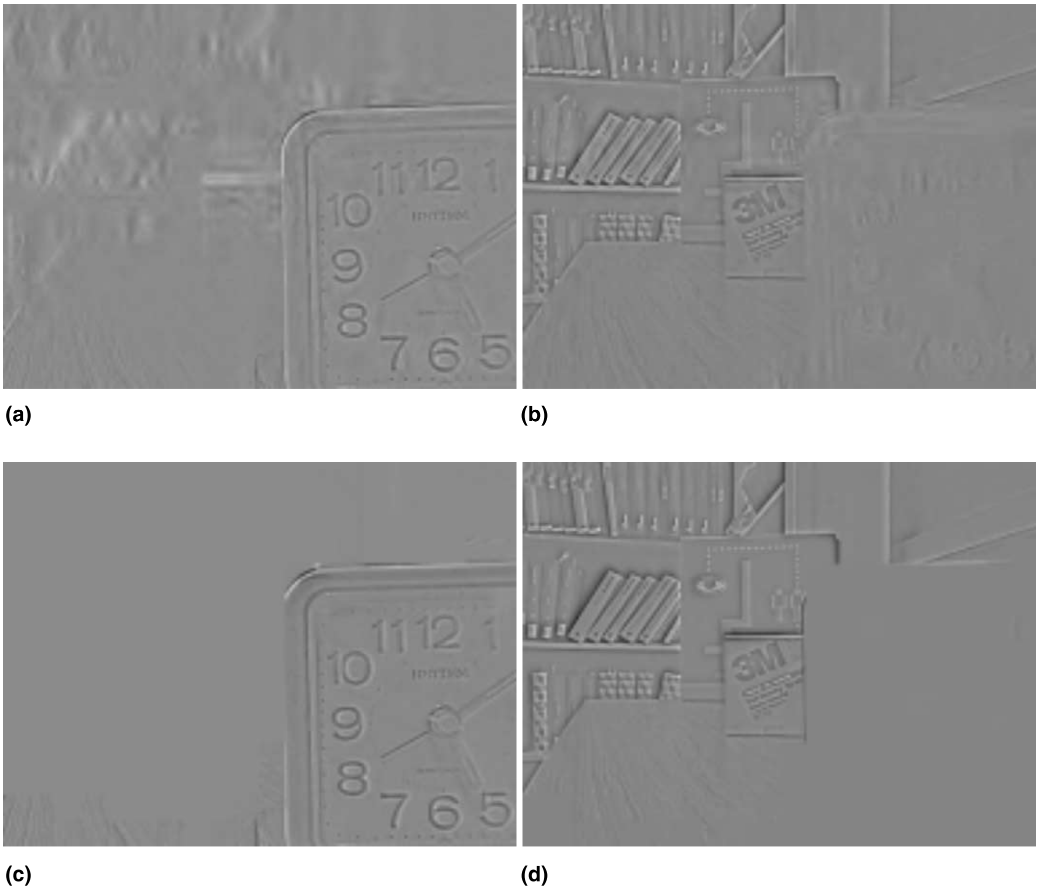

Fig. 11. Differences between the fused images in Fig. 10(c), (d) and source images in Fig. 10(a), (b). (a) Difference between the fused image using DWT (Fig. 10(c)) and source image Fig. 10(a). (b) Difference between the fused image using DWT (Fig. 10(c)) and source image Fig. 10(b). (c) Difference between the fused image using PNN (Fig. 10(d)) and source image Fig. 10(a). (d) Difference between the fused image using PNN (Fig. 10(d)) and source image Fig. 10(b).

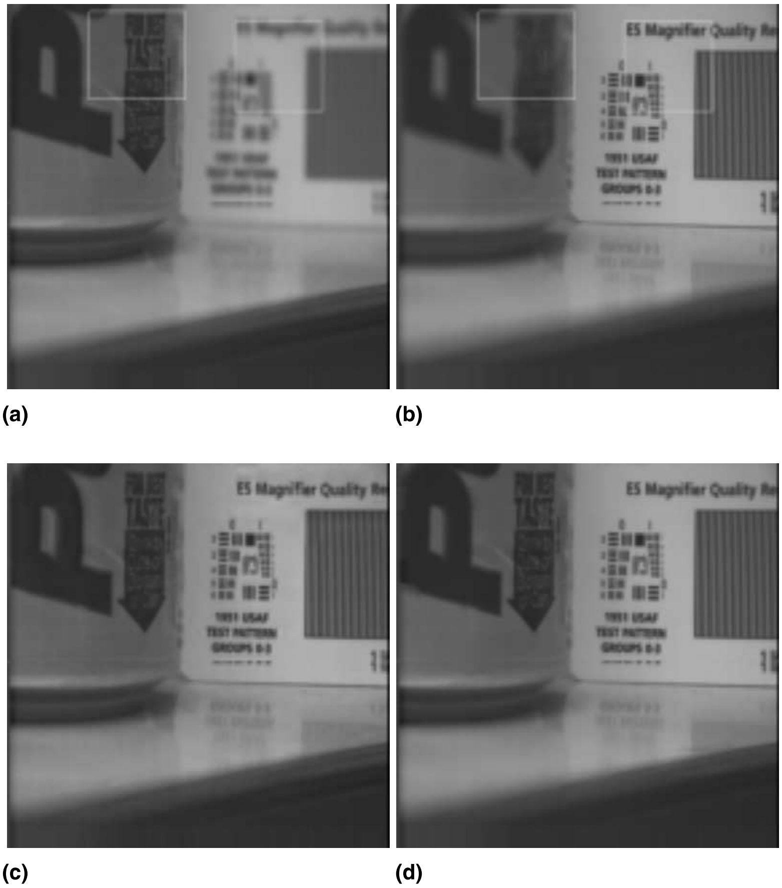

Fig. 12. The "Pepsi" source images and fusion results. The training set is selected from regions marked by the rectangles in (a) and (b). (a) Focus on the Pepsi can; (b) focus on the testing card; (c) fused image using DWT (coif5, level = 5) and (d) fused image using PNN.

clearer than the second image, and the reverse is true for the remaining 15 pairs. A training set with a total of 60 image block pairs is thus formed. The three features, SF, VI and EG, are extracted (with $\alpha = 0.6$ in (2)) and normalized to the range [0, 1] before feeding into the PNN and RBFN. Both neural networks contain three input units and one output unit. For the PNN, we have 60 hidden units [1] and $\sigma = 0.09$ in (3), whereas for the RBFN, we use 16 hidden units and $\sigma = 0.25$.

For comparison purposes, we also perform fusion using the DWT-based method mentioned in Section 1. The wavelet basis "coif5", together with a decomposition level of 5, is used. Similar to (Li et al., 1995), we employ a region-based activity measurement for the active level of the decomposed wavelet coefficients, a maximum selection rule for coefficient combination, together with a window-based consistency verification scheme.

Two evaluative criteria are used. They are the root mean squared error (RMSE)

$$RMSE = \sqrt{\frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}(R(i,j) - Z(i,j))^2},$$

between the reference image R and the fused image Z, and also the corresponding mutual information (MI)

$$MI = \sum_{i_1=1}^{L}\sum_{i_2=2}^{L} h_{R,Z}(i_1,i_2)\log_2\frac{h_{R,Z}(i_1,i_2)}{h_R(i_1)h_Z(i_2)}.$$

[1] Recall that the number of hidden units in a PNN is equal to the number of training patterns.

Here, $h_{R,Z}$ is the normalized joint gray level histogram of images R and Z, h_R, h_Z are the normalized marginal histograms of the two images, and L is the number of gray levels. Notice that *MI* measures the reduction in uncertainty about the reference image due to the knowledge of the fused image, and so a larger *MI* is preferred.

3.2.2. Results

The resultant fused images are shown in Fig. 6(d)–(f). A clearer comparison can be made by examining the differences between the fused and source images (Fig. 7). Recall that the focus in Fig. 6(b) is on the left while that in Fig. 6(c) is on the right. It can be seen from Fig. 7(b) that the fused image produced by PNN is basically a combination of the good-focus parts in the source images. Result for RBFN is very similar and so will not be reported here.

Quantitative comparison of their performance is shown in Table 3. Both neural networks again show significant improvement over the DWT-based method.

3.3. Subjective evaluation of the fusion method

3.3.1. Experimental setup

The second experiment is performed on the three sets of source images in Figs. 8, 10 and 12. Their sizes are 480 × 640, 480 × 640 and

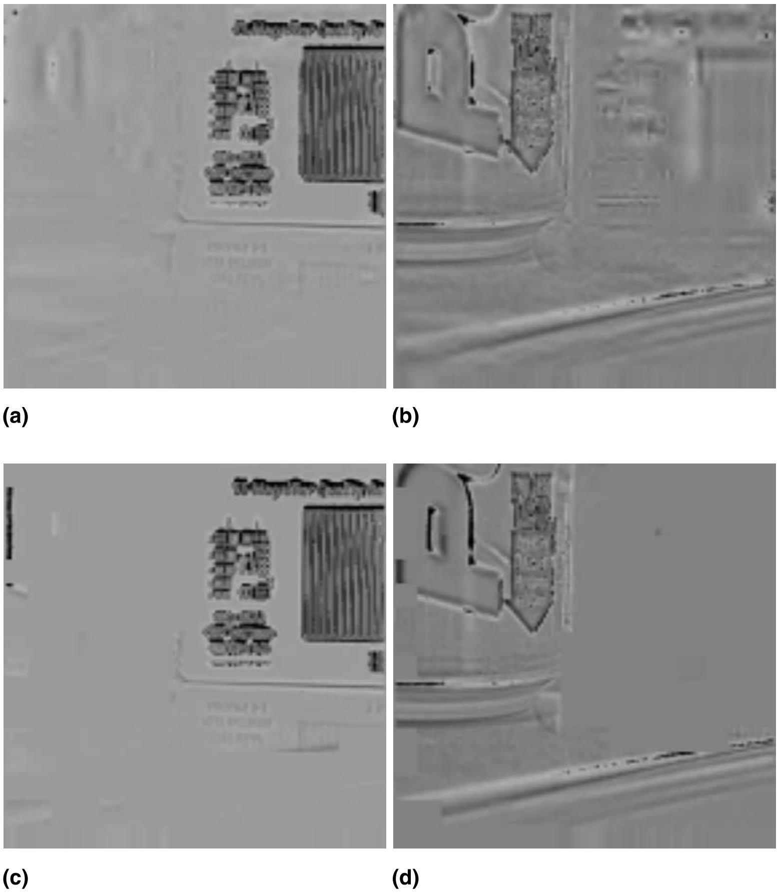

Fig. 13. Differences between the fused images in Fig. 12(c), (d) and source images in Fig. 12(a), (b). (a) Difference between the fused image using DWT (Fig. 12(c)) and source image Fig. 12(a). (b) Difference between the fused image using DWT (Fig. 12(c)) and source image Fig. 12(b). (c) Difference between the fused image using PNN (Fig. 12(d)) and source image Fig. 12(a). (d) Difference between the fused image using PNN (Fig. 12(d)) and source image Fig. 12(b).

Table 3
Performance of the different fusion methods on processing Fig. 6(b) and (c)

	DWT	PNN	RBFN
RMSE	1.5342	0.2634	0.2580
MI	6.2200	7.3940	7.2289

512 × 512, respectively. Here, each image contains multiple objects at different distances from the camera. Thus, one or more objects naturally become(s) out-of-focus when the image is taken. For example, the focus in Fig. 8(a) is on the clock, while that in Fig. 8(b) is on the student. The true gray value of each pixel is, however, not available and so only a subjective visual comparison is intended here. As in Section 3.2.1, we use an image block size of 32 × 32 and a training set size of 60. The PNN and RBFN also use the same network topologies as in Section 3.2.1.

3.3.2. Results

Fusion results on using DWT and PNN are shown in Fig. 8(c), (d), Fig. 10(c), (d) and Fig. 12(c), (d). The RBFN produces very similar results as the PNN and so will not be reported here. Again, a clearer comparison of their performance can be made by examining the differences between the fused images and each source image (Figs. 9, 11 and 13). Take the "Lab" images as an example. Recall that the focus in Fig. 8(a) is on the clock while that in Fig. 8(b) is on the student. It can be seen from Fig. 9(c) and (d) that the fused image produced by PNN is basically a combination of the good-focus clock and the good-focus student. In comparison, the result by DWT is much inferior. In particular, notice that there is a slight movement of the student's head in Fig. 8, and this causes a lot of differences between the fused image and the good-focus student (Fig. 9(b)). Similar observations can be made for the other two sets of images.

4. Conclusion

In this paper, we combine the idea of image blocks and artificial neural networks for pixel level multifocus image fusion. Features indicating the clarity of an image block are extracted and fed into the neural network, which then learns to determine which source image is clearer at that particular physical location. Two neural network models, namely the PNN and RBFN, have been used. Experimental results show that this method outperforms the DWT-based approach, particularly when there is object movement or registration problems in the source images.

An issue that will be investigated in the future is on adaptive methods for choosing the image block size. In general, if the block size is too large, a particular block may contain two or more objects at different distances from the camera, and consequently will lead to a less clear image. On the other hand, using a very small block size may lead to the saw-tooth effect. Nevertheless, preliminary experiments suggest that the fusion result is not quite sensitive to this parameter, provided that its value is not too extreme.

References

Bellman, R.E., 1961. Adaptive Control Processes. Princeton University Press, Princeton, NJ.

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Oxford University Press, Oxford.

Burt, P.T., Andelson, E.H., 1983. The Laplacian pyramid as a compact image code. IEEE Trans. Comm. 31, 532–540.

Burt, P.J., Kolczynski, R.J., 1993. Enhanced image capture through fusion. In: Proc. 4th Internat. Conf. on Computer Vision, Berlin, Germany, pp. 173–182.

Canny, J., 1986. A computational approach to edge detection. IEEE Trans. Pattern Recognition Machine Anal. 8 (6), 679–698.

Chipman, L.J., Orr, Y.M., Graham, L.N., 1995. Wavelets and image fusion. In: Proc. Internat. Conf. on Image Processing, Washington, USA, pp. 248–251.

Eskicioglu, A.M., Fisher, P.S., 1995. Image quality measures and their performance. IEEE Trans. Comm. 43 (12), 2959–2965.

Hertz, J., Krogh, A., Palmer, R.G., 1991. Introduction to the Theory of Neural Computation. Addison-Wesley, Redwood City, CA.

Huang, J.W., Shi, Y.Q., Dai, X.H., 1999. A segmentation-based image coding algorithm using the features of human vision system. J. Image Graphics 4 (5), 400–404.

Koren, I., Laine, A., Taylor, F., 1995. Image fusion using steerable dyadic wavelet. In: Proc. Internat. Conf. on Image Processing, Washington, USA, pp. 232–235.

Li, H., Manjunath, B.S., Mitra, S.K., 1995. Multisensor image fusion using the wavelet transform. Graphical Models Image Processing 57 (3), 235–245.

Matsopoulos, G.K., Marshall, S., Brunt, J.N.H., 1994. Multiresolution morphological fusion of MR and CT images of the human brain. Proc. IEE: Vision, Image Signal Processing 141 (3), 137–142.

Moody, J., Darken, C., 1998. Learning with localized receptive fields. In: Proc. 1988 Connectionist Models Summer School, pp. 133–143.

Seales, W.B., Dutta, S., 1996. Everywhere-in-focus image fusion using controllable cameras. Proc. SPIE 2905, 227–234.

Specht, D.F., 1990. Probabilistic neural networks. Neural Networks 3, 110–118.

Toet, A., van Ruyven, L.J., Valeton, J.M., 1989. Merging thermal and visual images by a contrast pyramid. Opt. Engrg. 28 (7), 789–792.

Unser, M., 1995. Texture classification and segmentation using wavelet frames. IEEE Trans. Image Processing 4 (11), 1549–1560.

Yocky, D.A., 1995. Image merging and data fusion by means of the discrete two-dimensional wavelet transform. J. Opt. Soc. Am. A: Opt., Image Sci. Vision 12 (9), 1834–1841.

Zhang, Z., Blum, R.S., 1999. A categorization of multiscale-decomposition-based image fusion schemes with a performance study for a digital camera application. Proc. IEEE 87 (8), 1315–1326.