An Efficient Deep Learning Framework for Fusing Multimodal Images in Computer Vision Tasks

G.VENKATA KAVYA[1], SK.MUNWAR ALI[2]

[1]PG Scholar, Dept. of E.C.E, Eswar College Of Engineering, Narasaraopet, Palnadu, Dt.

[2]Asst Professors, Dept. of E.C.E, Eswar College Of Engineering, Narasaraopet, Palnadu, Dt.

In medical imaging, techniques like CT, MRI, PET, and SPECT offer complementary structural and functional information. Image fusion aims to integrate this diverse data to enhance contrast, clarity, and clinical usability. An effective fusion method should preserve all relevant source image details without introducing artefacts or misalignment. The proposed approach uses a Siamese CNN model with identical weight branches for consistent feature extraction, improving training efficiency and fusion quality. It integrates Gaussian pyramid decomposition and multi scale transformation for perceptually aligned fusion, while a localized similarity-based strategy adaptively refines coefficients. Gaussian pyramid decomposition is used, and the pyramid transform is used for multi scale decomposition, so that the fusion process is more in line with human visual perception. In addition, the localized similarity-based fusion strategy is used to adaptively adjust the decomposed coefficients. And CNN-driven approaches, delivering superior fused images for medical diagnosis.

Key Words: Deep Learning, Fusion Method, Neural Network, Medical Imaging, Convolutional Network, Image Quality

1. Introduction

With the rapid development of sensor and computer technology, medical imaging has emerged as an irreplaceable component in various clinical applications including diagnosis, treatment planning and surgical navigation. To provide medical practitioners sufficient information for clinical purposes, medical images obtained with multiple modalities are usually required, such as X-ray, computed tomography (CT), magnetic resonance (MR), positron emission tomography (PET), single photon emission computed tomography (SPECT), etc. Due to the difference in imaging mechanism, medical images with different modalities focus on different categories of organ/tissue information. For instance, the CT images are commonly used for the precise localization of dense structures like bones and implants, the MR images can provide excellent soft-tissue details with high-resolution anatomical information, while the functional information on blood flow and metabolic changes can be offered by

Doi:10.48047/ijiee.2025.15.6.16

PET and SPECT images but with low spatial resolution. Multi-modal medical image fusion aims at combining the complementary information contained in different source images by generating a composite image for visualization, which can help physicians make easier and better decisions for various purposes

Recent advancements in medical image fusion have focused on multi-scale transform (MST) techniques to handle the intensity variations caused by differing imaging mechanisms. Traditional MST-based fusion involves decomposition, fusion, and reconstruction steps using transforms like pyramids, wavelets, contour lets, and shear lets. A key challenge in this process is creating accurate weight maps through activity level measurement and weight assignment, which often suffer from issues like noise, misregistration, and intensity differences. To address these limitations, this study introduces a convolutional neural network (CNN) that learns an end-to-end mapping from source images to weight maps, integrating both steps optimally. The method also uses image pyramids for perceptual alignment and a local similarity-based strategy to adaptively refine fusion, improving robustness and visual quality.

2. Related Work

In our recent work [8], a CNN-based multi-focus image fusion method which can obtain state-of-the-art results was proposed. In the method, two source images are fed to the two branches of a Siamese convolutional network in which the two branches share the same architecture and weights respectively. Each branch contains three convolutional layers and the obtained feature maps essentially act as the role of activity level measures. Issue, we adopt a local similarity-based fusion strategy to determine the fusion mode for the decomposed coefficients [12]. When the contents of source images have high similarity, the “weighted-average” fusion mode is applied to avoid losing useful information. In this situation, the weights obtained by the CNN are more reliable than the coefficient-based measure, so they are employed as the merging weights. When the similarity of image contents is low, the “choose-max” or “selection” fusion mode is preferred to mostly preserve the salient details from source images. In this situation, the CNN output is not reliable, and the pixel activity is

Doi:10.48047/ijiee.2025.15.6.16

directly measured by the absolute values of the decomposed coefficients.

3. Existing System

The current image fusion methods are divided into two categories. One is to directly fuse source images in spatial domain. However, this kind of methods is not good in dealing with edge. The other one is to integrate source images in transform domain. This type of approaches could remove the block effect and get more consistent fusion result. Image fusion methods based on MSD draw researcher's attention in recent years. For example, Discrete wavelet transform based method[8, 9], stationary wavelet transform based method[10], non-sub sampled contour let transform based method DWT, as a popular MSD tool, is first proposed, It provides a richer scale space analysis for image compared to other MSD tools because it can decompose image into magnitude and phase information. The magnitude of DWT is near shift invariant so it have better texture representation than wavelet and complex wavelet, and the phase of DWT contains richer geometric information.

3.1 Standard Image Fusion Methods

Image fusion methods can be broadly classified into two - spatial domain fusion and transform domain fusion. Another important spatial domain fusion method is the high pass filtering based technique. Here the high frequency details are injected into up sampled version of MS images. The disadvantage of spatial domain approaches is that they produce spatial distortion in the fused image. Spectral distortion becomes a negative factor while we go for further processing, such as classification problem distortion can be very well handled by transform domain approaches on image fusion. The multi resolution analysis has become a very useful tool for analysing remote sensing images.

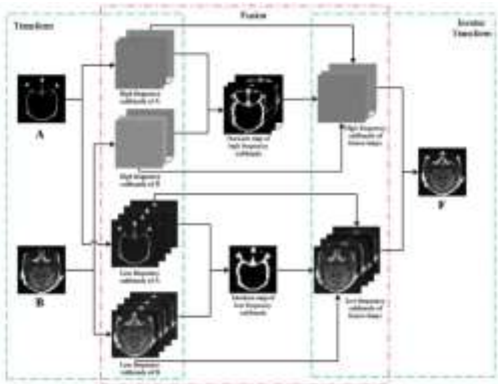


Fig 1: Standard Image Fusion

The Discrete Wavelet Transform (DWT) is widely used in image fusion due to its ability to separate images into different frequency components. While low-pass filtering reduces the image resolution by removing high-frequency details, it does not change the image scale; sub-sampling

Doi:10.48047/ijiee.2025.15.6.16

afterward doubles the scale by discarding redundant samples without losing information. Most of the image's energy resides in the low-frequency part, which forms the approximation of the fused image and strongly influences its overall perception. This paper proposes a weighted average fusion rule for the low-frequency sub bands, using a comprehensive feature that combines the phase, magnitude, and spatial variance of the low-frequency coefficients. The low-frequency DWT components are represented by one magnitude matrix and three phase matrices, capturing local shifts and texture information to improve fusion quality. where α is a visual constant and is determined by a physiological visual test in a range of 0.6 to 0.7, L(i,j) and H(i,j) are low frequency coefficient and high frequency coefficient, respectively. The relationship between contrast and background intensity is non-linear, which makes the human visual system highly sensitive to contrast variations. To provide further quantitative comparison of various fusion methods, objective evaluations of fused results are given in this section. The spatial frequency (SF) and MSE ,PSNR are adopted as evaluation indexes. Large values indicate a better fusion result for all indexes

4. Proposed Method

The convolutional network used in the proposed fusion algorithm. It is a Siamese network in which the weights of the two branches are constrained to the same. Each branch consists of three convolutional layers and one max-pooling layer which is the same as the network used in [24]. To reduce the memory consumption as well as increase the computational efficiency, we adopt a much slighter model in this work by removing a fully connected layer from the network used in [24]. The 512 feature maps after concatenation are directly connected to a 2-dimensional vector. It can be calculated that the slight mode only takes up about 1.66 MB of physical memory in single precision, which is significantly less than the 33.6 MB model employed in [14]. Finally, this 2dimensional vector is fed to a 2-way Soft Max layer (not shown in Fig.2), which produces a probability distribution over two classes. The two classes correspond two kinds of normalized weight assignment results, namely, "first patch 1 and second patch 0" and "first patch 0 and second patch 1", respectively. The probability of each class indicates the

Doi:10.48047/ijiee.2025.15.6.16

possibility of each weight assignment. In this situation, also considering that the sum of two output probabilities is 1, the probability of each class just indicates the weight assigned to its corresponding input patch.

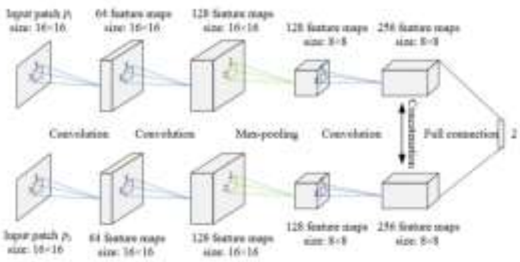
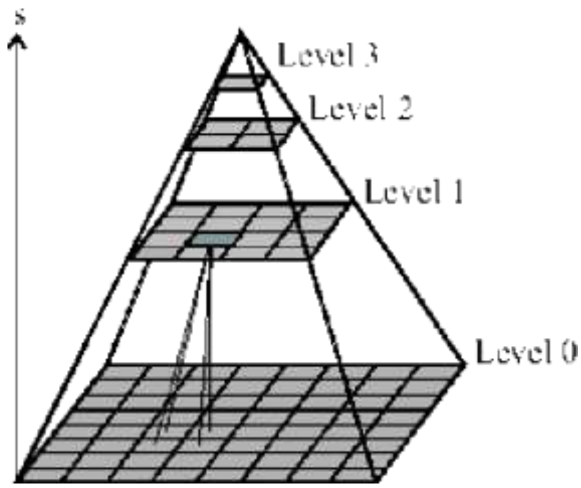


Fig.2: Architecture for CNN training

4.1 Pyramid Generation

There are two main types of pyramids: low pass and band pass. A low pass pyramid is made by smoothing the image with an appropriate smoothing filter and then subsampling the smoothed image, usually by a factor of 2 along each coordinate direction. The resulting image is then subjected to the same procedure, and the cycle is repeated multiple times. Each cycle of this process results in a smaller image with increased smoothing, but with decreased spatial sampling density (that is, decreased image resolution). If illustrated graphically, the entire multi-scale representation will look like a pyramid, with the original image on the bottom and each cycle's resulting smaller image stacked one atop the other. A band pass pyramid is made by forming the difference between images at adjacent levels in the pyramid and performing image interpolation between adjacent levels of resolution, to enable computation of pixel wise differences.

The Gaussian pyramid is computed as follows. The original image is convolved with a Gaussian kernel. As described above the resulting image is a low pass filtered version of the original image. The cut-off frequency can be controlled using the parameter σ. The Laplacian is then computed as the difference between the original image and the low pass filtered image. This process is continued to obtain a set of band-pass filtered images (since each is the difference between two levels of the Gaussian pyramid). Thus, the Laplacian pyramid is a set of band pass filters



Doi:10.48047/ijiee.2025.15.6.16

Fig 3: the filtered images stacked one on top of the other form a tapering pyramid structure, hence the name

Step 1: CNN-based weight map generation. Feed the two source images A and B to the two branches of the convolutional network, respectively. The weight map W is generated using the approach described above.

Step 2: Pyramid decomposition. Decompose each source image into a Laplacian pyramid. Let $\{A\}$ and $\{B\}$ respectively denote the pyramids of A and B, where l indicates the l-th decomposition level. Decompose the weight map W into a Gaussian pyramid $\{W\}$. The total decomposition level of each pyramid is set to the highest possible value⌊log2 min(H,W)⌋, where H×W is the spatial size of source images and⌊·⌋denotes the flooring operation.

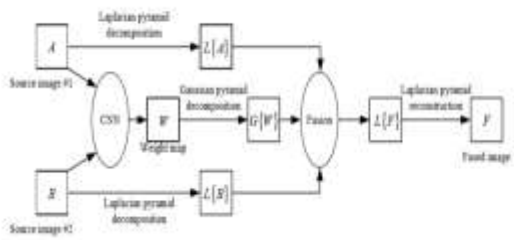


Fig.4: Proposed medical image fusion

The inverse transform to obtain the original image g_0 from the N detail images L_0, L_1, …, L_N and the low pass image g_N is as follows:

1. g_N is up sampled by inserting zeros between the sample values and interpolating the missing values by convolving it with the filter w to obtain the image g'_N.
2. The image g'_N is added to the lowest level detail image L_N to obtain the approximation image at the next upper level:

$$g_{N-1} = L_N + g'_N$$

3. Steps 1 and 2 are repeated on the detail images L_0, L_1, …, L_{N-1} to obtain the original image.

The multimodal image fusion method based on deep learning utilizes a pyramid-based CNN architecture to effectively combine information from different medical imaging modalities, such as CT and MRI. This approach decomposes source images into multi-scale representations using image pyramids, capturing features at various levels of detail. A convolutional neural network (CNN) is then trained to generate weight maps that guide the fusion process, allowing it to jointly perform activity level measurement and weight assignment in an optimized manner. By integrating local similarity strategies and multi-scale decomposition, the method enhances contrast, reduces artifacts, and ensures that

Doi:10.48047/ijiee.2025.15.6.16

the fused image retains complementary information from all input sources, resulting in improved diagnostic quality.

5 .Results and Discussion

The experiments were conducted using MATLAB 2016b on a high-speed CPU to ensure faster execution, using the test images shown in Fig5. The goal of any image fusion algorithm is to effectively combine essential information from both source images into a single output image. The quality of the fused image is assessed both visually and quantitatively using fusion metrics to provide a comprehensive evaluation of the proposed and existing methods. Fused image cannot be judged exclusively by seeing the output image or by measuring fusion metrics. It should be judged qualitatively using visual display and quantitatively using fusion metrics. In this section, we are presenting both visual quality and quantitative analysis of proposed and existing algorithms such as,

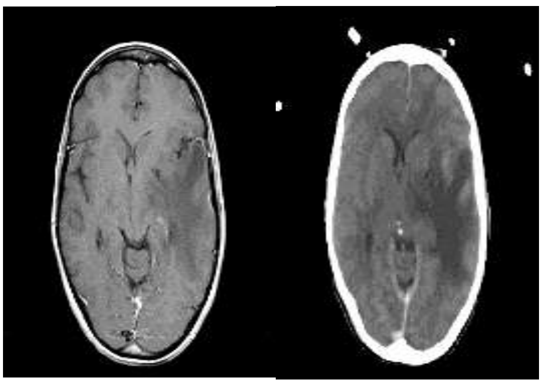
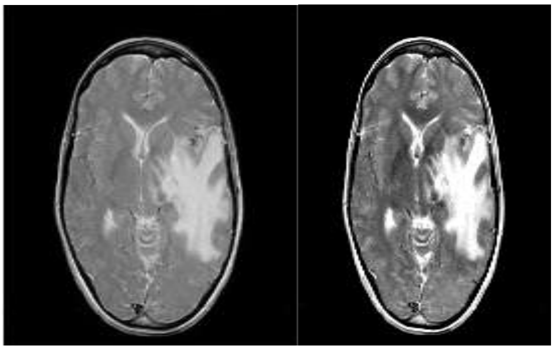


Fig 5: Test images (a) dataset (MR)& (b) dataset (CT)



(a) (b)

Fig 6: Obtained fused images for dataset using (a) DWT (b) Proposed method

However, all the existing fusion methods outputs not good at visual perception, lack of contrast with edge information and texture preservation. Our proposed method with datasets Fig 6 Obtained fused images for dataset using (a) DWT (b) Proposed method which are presented in looks more quality in visualization, good contrast with proper edge information and excellent texture preservation as the value of entropy is much higher.

6. Conclusion & Future work

In this paper, a medical image fusion method based on convolutional neural networks is proposed. We employ a Siamese network to generate a direct mapping from source images to a weight map which contains the integrated pixel activity information. The main novelty of this approach is it can jointly implement

Doi:10.48047/ijiee.2025.15.6.16

activity level measurement and weight assignment via network learning, which can overcome the difficulty of artificial design. To achieve perceptually good results, some popular techniques in image fusion such as multi-scale processing and adaptive fusion mode selection are appropriately adopted. Experimental results demonstrate that the proposed method can obtain high-quality results in terms of visual quality and objective metrics. In addition to the proposed algorithm itself, another contribution of this work is that it exhibits the great potential of some deep learning techniques for image fusion, which will be further studied in the future.

REFERENCES

[1] A. James and B. Dasarathy, “Medical image fusion: a survey of the state of the art,” Information Fusion, vol. 19, pp. 4–19, 2014.

[2] L. Yang, B. Guo, and W. Ni, “Multimodality medical image fusion based on multi scale geometric analysis of contourlet transform,” Neurocomputing, vol. 72, pp. 203–211, 2008.

[3] Z. Wang and Y. Ma, “Medical image fusion using m-pcnn,” Information Fusion, vol. 9, pp. 176–185, 2008.

[4] B. Yang and S. Li, “Pixel-level image fusion with simultaneous orthogonal matching pursuit,” Information Fusion, vol. 13, pp. 10–19, 2012.

[5] S. Li, H. Yin, and L. Fang, “Group-sparse representation with dictionary learning for medical image denoising and fusion,” IEEE Transactions on Biomedical Engineering, vol. 59, pp. 3450–3459, 2012.

[6] Z. L. G. Bhatnagar, Q. Wu, “Directive contrast based multimodal medical image fusion in nsct domain,” IEEE Transactions on Multimedia, vol. 15, pp. 1014–1024, 2013.

[7] S. Li, X. Kang, and J. Hu, “Image fusion with guided filtering,” IEEE Transactions on Image Processing, vol. 22, no. 7, pp. 2864–2875, 2013.

[8] R. Shen, I. Cheng, and A. Basu, “Cross-scale coefficient selection for volumetric medical image fusion,” IEEE Transactions on Biomedical Engineering, vol. 60, pp. 1069–1079, 2013.

[9] R. Singh and A. Khare, “Fusion of multimodal medical images using daubechies complex wavelet transform c a multiresolution approach,” Information Fusion, vol. 19, pp. 49–60, 2014.

[10] L. Wang, B. Li, and L. Tan, “Multimodal medical volumetric data

Doi:10.48047/ijiee.2025.15.6.16

fusion using 3-d discrete shearlet transform and global-to-local rule," IEEE Transactions on Biomedical Engineering, vol. 61, pp. 197–206, 2014.

[11] Z. Liu, H. Yin, Y. Chai, and S. Yang, "A novel approach for multimodal medical image fusion," Expert Systems with Applications, vol. 41, pp. 7425–7435, 2014.

[12] G. Bhatnagar, Q. Wu, and Z. Liu, "A new contrast based multimodal medical image fusion framework," Neurocomputing, vol. 157, pp. 143– 152, 2015.

[13] Y. Liu, S. Liu, and Z. Wang, "A general framework for image fusion based on multi-scale transform and sparse representation," Information Fusion, vol. 24, no. 1, pp. 147–164, 2015.

[14] Q. Wang, S. Li, H. Qin, and A. Hao, "Robust multi-modal medical image fusion via anisotropic heat diffusion guided low-rank structural analysis," Information Fusion, vol. 26, pp. 103–121, 2015.

[15] Y. Liu and Z. Wang, "Simultaneous image fusion and denosing with adaptive sparse representation," IET Image Process., vol. 9, no. 5, pp. 347–357, 2015.

[16] Y. Yang, Y. Que, S. Huang, and P. Lin, "Multimodal sensor medical image fusion based on type-2 fuzzy logic in nsct domain," IEEE Sensors Journal, vol. 16, pp. 3735–3745, 2016.

[17] J. Du, W. Li, B. Xiao, and Q. Nawaz, "Union laplacian pyramid with multiple features for medical image fusion," Neuro computing, vol. 194, pp. 326–339, 2016.

Doi:10.48047/ijiee.2025.15.6.16