Image Fusion using Deep Learning and Traditional Methods

Rishit Puri
Dept of Electronics and Communication Engineering
SRM Institute of Science and Technology
SRM University, Kattankulathur
Chennai, India

Kavya Shruthi
Dept of Electronics and Communication Engineering
SRM Institute of Science and Technology
SRM University, Kattankulathur
Chennai, India

***Abstract*: This paper is a survey paper based on the topic of image fusion using deep learning. The paper talks about the traditional image fusion and image decomposition methods along with deep learning methods for image decomposition. Image fusion aims at obtaining fused image keeping the advantage of source images. Traditional image decomposition includes methods such as multi-scale decomposition, two scale decomposition etc. Traditional image fusion techniques include DWT, DCT, Top Hat etc.**

Keywords—Image fusion, image decomposition, Deep Learning, infrared image, visible image.

I. INTRODUCTION

Image fusion is an important task in image processing. It aims to extract important features from images of multi-modal sources and reconstruct the fused image using the complementary information conveyed by the multiple pictures by means of a fusion strategy. Image fusion has numerous applications such as autonomous driving, medical imaging, defogging, security surveillance etc.

At present there are two types of image decomposition available which are traditional methods and the deep learning methods. Few of the traditional methods are multiscale decomposition, two scale decomposition etc. In this paper multiscale and two scale decomposition is explained briefly and in the later part Deep learning technique is explained.

A. Multi-Scale Decomposition

During the image fusion process, an image is typically decomposed into its detail and background layers using a technique called multi-scale decomposition. This involves applying a series of low-pass and high-pass filters to the image at different scales, which separates the low-frequency background information from the high-frequency detail information. The filtered images are then recombined to form the fused image, which preserves both the background and detail information from the original image. The specific method for image decomposition can vary depending on the application and desired outcome.

B. Two-Scale Decomposition

Two-scale decomposition is a method used to decompose an image into two layers: a coarse approximation of the image and a detail layer that contains the high-frequency information. The decomposition is done in two scales, hence the name two-scale decomposition.

The decomposition process typically begins by applying a low-pass filter to the image, which is used to create the coarse approximation layer. This layer contains the low-frequency information and serves as a rough representation of the image. Next, the difference between the original image and the coarse approximation layer is computed, which results in the detail layer. This layer contains the high-frequency information and represents the fine details of the image.

The two-scale decomposition can be done using various techniques, such as wavelet transform, curvelet transform and contourlet transform.

Two-scale decomposition is useful in image processing and computer vision applications, such as image compression, enhancement, and fusion. The two-scale decomposition can be used as a preprocessing step for image fusion, where a fused image can be created by combining the coarse approximation and detail layers from multiple images.

II. LITERATURE SURVEY

One mathematical equation often used in multi-scale decomposition is the wavelet transform. The wavelet transform decomposes a signal into different frequency components, represented by a set of coefficients. The general equation for the discrete wavelet transforms (DWT) of a signal x(n) is as follows:

$x(n) = \sum h(k)y(n-k) + \sum g(k)w(n-k)$

where h(k) and g(k) are the wavelet and scaling functions respectively, y(n) and w(n) are the approximation and detail coefficients at different levels of decomposition.

Another common equation used in multi-scale decomposition is the Laplacian Pyramid, which is a technique that recursively applies a low-pass filter to an image and then subtracts the filtered image from the original image to obtain a set of detail images at different scales. The general equation for the Laplacian Pyramid is:

$L(i,j) = G(i,j) - \text{Expand}[G(i,j)]$

where G(i,j) is the filtered image at a certain scale, L(i,j) is the Laplacian pyramid at that scale and Expand is the operation of upsampling and applying a low pass filter to an image. Both DWT and Laplacian pyramid are widely used in image processing for multi-scale decomposition.

One popular approach for image decomposition using deep learning is using Variational Autoencoders (VAEs). VAEs consist of an encoder network that maps the input image to a low-dimensional latent space, and a decoder network that maps the latent space back to the original image space. The encoder and decoder networks are trained jointly to reconstruct the input image. By considering the lowest level of the encoder as a bottleneck, the VAE can force the encoder to learn the most salient features of the image, which can be used for image decomposition.

Another approach is using Generative Adversarial Networks (GANs) which consist of two networks, a generator network and a discriminator network. The generator network generates new images from a random noise, while the discriminator network is trained to differentiate between real and generated images. By training the generator network to generate images similar to the input image, the generator can learn to decompose the input image into its constituent parts, such as texture, shape, and color. In summary, deep learning-based image decomposition methods use neural networks to decompose an image into its constituent parts or features, this can be done using various architectures such as autoencoders or generative models like VAEs and GANs.

This paper proposes a novel auto-encoder (AE) based fusion network. The core idea is that the encoder decomposes an image into background and detail feature maps with low- and high-frequency information, respectively, and that the decoder re-covers the original image. To this end, the loss function makes the background/detail feature maps of source images similar/dissimilar. In the test phase, background and detail feature maps are respectively merged via a fusion module, and the fused image is recovered by the decoder.

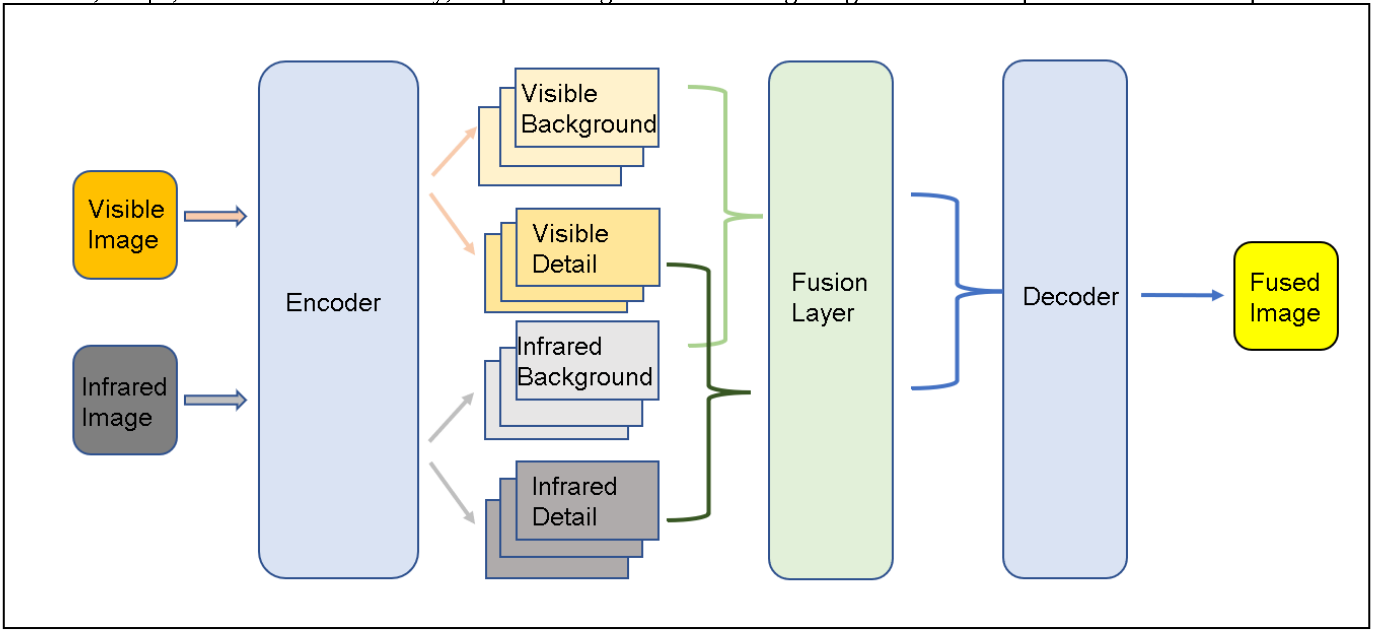


Image decomposition using autoencoders and neural networks involves training a neural network, called an autoencoder, to decompose an input image into its constituent parts or features. The autoencoder is trained to reconstruct the input image from its decomposed parts, and during the training process, it learns to extract the important features of the image that are necessary for reconstruction. The architecture of an autoencoder typically consists of an encoder and a decoder. The encoder is a neural network that maps the input image to a lower-dimensional feature space, called the bottleneck or latent space. The decoder is another neural network that maps the feature space back to the original image space. The encoder and decoder are trained together to reconstruct the input image as accurately as possible.

During the training process, the autoencoder is presented with a set of images and it tries to learn a mapping between the input image and its corresponding output image. The output image is obtained by passing the input image through the encoder and then through the decoder. The autoencoder learns to minimize the difference between the input image and the output image, and this forces the encoder to learn the most salient features of the image that are necessary for reconstruction. After training, the encoder can be used to decompose an input image into its constituent parts or features by passing it through the encoder and getting the feature space. The feature space can be visualized as a set of feature maps, each representing a different aspect of the image, such as texture, shape, and color. These feature maps can be further used for various image processing tasks, such as image classification, segmentation, and generation. In summary, autoencoders are neural networks trained to reconstruct an input image, as a result of this process, the encoder learns to extract the most salient features of the image, which can be used for image decomposition.

The mathematical formula involved in image decomposition using autoencoders is based on the principles of neural networks and optimization. The goal of an autoencoder is to learn a mapping between the input image and its corresponding output image, while also learning a compressed representation of the image in the bottleneck or latent space.

The basic formula for an autoencoder can be written as:

$$x' = f(W\_enc * f(W\_dec * z))$$

Where x is the input image, x' is the output image, z is the bottleneck or latent space, W_enc and W_dec are the weights of the encoder and decoder networks respectively, and f( ) is the activation function.

The autoencoder is trained to minimize the difference between the input image and the output image, which is often measured using a loss function such as mean squared error (MSE) or binary cross-entropy (BCE). The loss function is used to evaluate the quality of the reconstruction and to guide the optimization process.

The optimization process is often done using an optimization algorithm such as stochastic gradient descent (SGD) or Adam, which updates the weights of the network to minimize the loss function.

In summary, the mathematical formula involved in image decomposition using autoencoders is based on the principles of neural networks and optimization, where the goal is to learn a mapping between the input image and its corresponding output image while also learning a compressed representation of the image in the bottleneck or latent space. The optimization process is done by minimizing the difference between the input and output images using a loss function and an optimization algorithm.

Loss Function

In the training phase, we aim to obtain an encoder that performs two-scale decomposition on the source images, and at the same time, acquire a decoder that can fuse the images and preserve the information of source images well.

Image decomposition.

Background feature maps are used to extract the common features of source images, while detail feature maps are used to capture the distinct characteristics from infrared and visible images. Therefore, we should make the gap of background feature maps small. In contrast, the gap of detail feature maps should be great. To this end, the loss function of image decomposition is defined as follow,

$$L_1 = \Phi\left(\|B_V - B_I\|_2^2\right) - \alpha_1 \Phi\left(\|D_V - D_I\|_2^2\right)$$

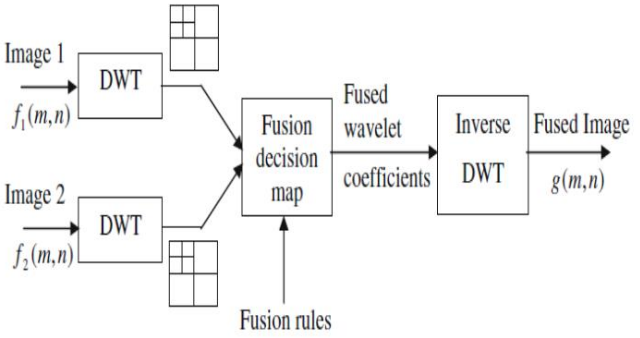
where BV, DV are the background and detail feature maps of the visible image V, and BI, DI are those of the infrared image I. Φ (·) is tanh function that is used to bound gap into interval (−1, 1).

Fusion Techniques

Image fusion using the discrete wavelet transform (DWT) is a process that combines multiple images into a single image by extracting features from each source image and then combining them in a way that preserves important information while reducing redundancy. The general process can be broken down into the following steps:

1. Transform the source images into the frequency domain using the DWT. This decomposes the images into different frequency bands, such as low-frequency and high-frequency bands, which represent the coarse and fine details of the image, respectively.
2. Extract the features from each source image by selecting the most important frequency bands. This can be done using various selection criteria, such as energy, variance, or entropy.
3. Combine the selected features from each source image to form a single fused image. This can be done using various fusion rules, such as averaging, maximum, or principal component analysis (PCA).
4. Transform the fused image back into the spatial domain using the inverse DWT (IDWT) to obtain the final fused image.

The specific details of the process, such as the type of wavelet function, the level of decomposition, and the fusion rule, can vary depending on the application and the desired outcome. However, the key idea behind image fusion using DWT is to extract the most important features from each source image and then combine them in a way that preserves important information while reducing redundancy.



The DWT of an image x(n) is computed by convolving the image with a set of wavelet functions, h(k) and g(k) , which are also called the analysis and synthesis functions, respectively. The general equation for the DWT is as follows:

x(n) = ∑h(k)y(n-k) + ∑g(k)w(n-k)

where y(n) and w(n) are the approximation and detail coefficients at different levels of decomposition. The inverse discrete wavelet transform (IDWT) is used to reconstruct the image from the wavelet coefficients. The IDWT is the reverse process of the DWT, and it can be used to obtain the final fused image from the wavelet coefficients. The general equation for the IDWT is:

x(n) = ∑h(k)y(n-k) + ∑g(k)w(n-k)

DCT

The mathematical formula involved in image fusion using the discrete cosine transform (DCT) is based on the principles of linear algebra, signal processing, and the DCT. The DCT is a mathematical technique used to decompose a signal into a sum of cosine functions of different frequencies. The Discrete Cosine Transform (DCT) of an image x(n) is computed by applying the DCT to each block of pixels in the image. The DCT of a block of pixels is represented by a matrix C, where each element c(u,v) is given by the following formula:

c(u,v) = α(u) * α(v) * ∑m=0^(N-1) ∑n=0^(N-1) x(m,n) * cos((2m+1)uπ/(2N)) * cos((2n+1)vπ/(2N))

Where x(m,n) is the value of the pixel at position (m,n) in the block, N is the size of the block, and α(u) and α(v) are normalization factors.

After the DCT is applied to each block of the image, it results in a set of frequency coefficients, where each coefficient represents a different spatial frequency component of the image. These coefficients can be used to extract the most important features from each source image, then they can be combined to form a single fused image using various fusion rules, such as averaging, maximum, or principal component analysis (PCA). The final fused image can be obtained by applying the inverse DCT (IDCT) to the fused coefficients.

REFERENCES

[1] [Bavirisetti and Dhuli, 2015] Durga Prasad Bavirisetti and Ravindra Dhuli. Fusion of infrared and visible sensor im- ages based on

anisotropic diffusion and karhunen-loeve transform. IEEE Sensors Journal, 16(1):203–209, 2015.

[2] [Bavirisetti and Dhuli, 2016] Durga Prasad Bavirisetti and Ravindra Dhuli. Two-scale image fusion of visible and infrared images using saliency detection. Infrared Physics & Technology, 76:52–64, 2016.

[3] [Bhatnagar and Liu, 2015] Gaurav Bhatnagar and Zheng Liu. A novel image fusion framework for night-vision naviga- tion and surveillance. Signal, Image and Video Processing, 9(1):165–175, 2015.

[4] [Brown and Susstrunk, 2011 ¨] Matthew Brown and Sabine Susstrunk. Multi-spectral sift for scene category recog- ¨ nition. In CVPR 2011, pages 177–184. IEEE, 2011.

[5] [Guo et al., 2017] Hanqi Guo, Yong Ma, Xiaoguang Mei, and Jiayi Ma. Infrared and visible image fusion based on total variation and augmented lagrangian. Journal of the Optical Society of America A, 34(11):1961–1968, 2017.[Hu et al., 2017]

[6] Hai-Miao Hu, Jiawei Wu, Bo Li, Qiang Guo, and Jin Zheng. An adaptive fusion algorithm for visible and infrared videos based on entropy and the cumulative distri- bution of gray levels. IEEE Transactions on Multimedia, 19(12):2706–2719, 2017.

[7] [Lahoud and Susstrunk, 2018] F. Lahoud and S. Susstrunk. Ar in vr: Simulating infrared augmented vision. In 2018 25th IEEE International Conference on Image Processing (ICIP), pages 3893–3897, Oct 2018.

[8] [Lahoud and Susstrunk, 2019 ¨] Fayez Lahoud and Sabine Susstrunk. Fast and efficient zero-learning image fusion. ¨ arXiv preprint arXiv:1905.03590, 2019.

[9] [Li and Wu, 2018] Hui Li and Xiao-Jun Wu. Densefuse: A fusion approach to infrared and visible images. IEEE Transactions on Image Processing, 28(5):2614–2623, 2018.

[10] [Li et al., 2011] Shutao Li, Bin Yang, and Jianwen Hu. Performance comparison of different multi-resolution transforms for image fusion. Information Fusion, 12(2):74–84, 2011.

[11] [Li et al., 2018] Hui Li, Xiao-Jun Wu, and Josef Kittler. Infrared and visible image fusion using a deep learning framework. In 2018 24th International Conference on Pattern Recognition (ICPR), pages 2705–2710. IEEE, 2018.

[12] [Liu et al., 2016] Yu Liu, Xun Chen, Rabab K Ward, and Z Jane Wang. Image fusion with convolutional sparse representation. IEEE Signal Processing Letters, 23(12):18821886, 2016.

[13] [Ma et al., 2016] Jiayi Ma, Chen Chen, Chang Li, and Jun Huang. Infrared and visible image fusion via gradient transfer and total variation minimization. Information Fusion, 31:100–109, 2016.

[14] [Ma et al., 2019a] Jiayi Ma, Yong Ma, and Chang Li. Infrared and visible image fusion methods and applications: A survey. Information Fusion, 45:153–178, 2019.

[15] [Ma et al., 2019b] Jiayi Ma, Wei Yu, Pengwei Liang, Chang Li, and Junjun Jiang. Fusiongan: A generative adversarial network for infrared and visible image fusion. Information Fusion, 48:11–26, 2019.

[16] [Ma et al., 2020] Jiayi Ma, Pengwei Liang, Wei Yu, Chen Chen, Xiaojie Guo, Jia Wu, and Junjun Jiang. Infrared and visible image fusion via detail preserving adversarial learning. Information Fusion, 54:85–98, 2020.

[17] [Mao et al., 2016] Xiaojiao Mao, Chunhua Shen, and Yu-Bin Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In Advances in neural information processing systems, pages 2802–2810, 2016.

[18] [Meher et al., 2019] Bikash Meher, Sanjay Agrawal, Rutu parna Panda, and Ajith Abraham. A survey on region based image fusion methods. Information Fusion, 48:119–132, 2019.

[19] [Mirza and Osindero, 2014] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784, 2014.

[20] [Patil and Mudengudi, 2011] Ujwala Patil and Uma Mudengudi. Image fusion using hierarchical pca. In 2011 International Conference on Image Information Processing, pages 1–6. IEEE, 2011.

[21] [Prabhakar et al., 2017] K Ram Prabhakar, V Sai Srikar, and R Venkatesh Babu. Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. In ICCV, pages 4724–4732, 2017.

[22] [Ronneberger et al., 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, pages 234–241. Springer, 2015.

[23] [Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[24] [Toet and Hogervorst, 2012] Alexander Toet and Maarten A. Hogervorst. Progress in color night vision. Optical Engineering, 51(1):1 – 20, 2012.

[25] [Wang et al., 2004] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4):600–612, 2004.

[26] [Zhang et al., 2017] Xiaoye Zhang, Yong Ma, Fan Fan, Ying Zhang, and Jun Huang. Infrared and visible image fusion via saliency analysis and local edge-preserving multi-scale decomposition. Journal of the Optical Society of America A, 34(8):1400–1410, 2017.

[27] [Zong and Qiu, 2017] Jing-jing Zong and Tian-shuang Qiu. Medical image fusion based on sparse representation of classified image patches. Biomedical Signal Processing and Control, 34:195–205, 2017.

[28] "Study of different image fusion algorithm," International Journal of Emerging Technology and Advanced Engineering, vol. 3, no. 5, pp. 288–291, 2013.

[29] D. Yalamanda and V. P. Bhushan, "Medical image fusion based on improved wavelet transform."

[30] K. Rani and R. Sharma, "Study of image fusion using discrete wavelet and multiwavelet transform."

[31] V. Naidu and J. Raol, "Pixel-level image fusion using wavelets and principal component analysis,"Defence Science Journal, vol. 58, no. 3, pp. 338–352, 2008.

[32] Wang, Wencheng and Chang, Faliang, "A multi-focus image fusion method based on Laplacian pyramid," Journal of Computers, 2011.

[33] Li, Shutao and Kwok, JT-Y and Tsang, Ivor W and Wang, Yaonan, "Fusing images with different focuses using support vector machines," Neural Networks, IEEE Transactions on. IEEE, 2004, pp. 1555–1561

[34] Liang, Junli and He, Yang and Liu, Ding and Zeng, Xianju, "Image fusion using higher order singular value decomposition," Image Processing, IEEE Transactions on, vol. 31, no. 12, pp. 2898–2909, 2012 2013.

[35] Li, Shutao and Kang, Xudong and Hu, Jianwen, "Image fusion with guided filtering,"IEEE transactions on image processing: a publication of the IEEE Signal Processing Society, IEEE, vol. 27, no. 2, pp.2864–2875, 2013.

[36] K. He, J. Sun, and X. Tang, "Guided image filtering," in Computer Vision–ECCV 2010. Springer, 2010, pp. 1–14.

[37] Gu, W. Li, M. Zhu, and M. Wang, "Local edge-preserving multiscale decomposition for high dynamic range image tone mapping," Image Processing, IEEE Transactions on, vol. 22, no. 1, pp. 70–79, 2013