Multi-Focus Image Fusion Technique

Francis Chambers*, Mohammed Ibrahim Shariff†, Sebastian Ouslis‡, Shizhen Li§
* † ‡ § Faculty of Engineering
University of Waterloo, Waterloo, Ontario N2L 3G1

***Abstract*—All-in-focus (AIF) images, created by multi-focus image fusion, produce great advantage for computer vision with clear representation of all objects in field from different distances. In this article, a joint decision algorithm from different fusion techniques to maximize their respective strength in solving the fusion decision is proposed. The algorithm's intuition and execution is explained in great detail. An objective comparison between the proposed algorithm and current state-of-the-art algorithms is carried out based on different metrics. The proposed algorithm is also tested on its robustness against different noises.**

***Index Terms*—Multi-focus image fusion, Laplacian pyramid, Sum of Modified Laplacian, Image Variance**

I. Introduction

Multi-focus image fusion has emerged as a growing and innovative area of research in the field of computer vision. Due to the inherent depth-of-field limitations of traditional imaging systems, cameras can only focus on one object and surroundings, with rest of the image being out-of-focus. To solve this issue, multi-focus image fusion blend multiple images, each focused on a different depth plane, into a single all-in-focus composite image.

Existing fusion methods can be generalized into three different approaches: transform domain approach, spatial domain approach, and machine learning approach. In a transform domain based method, the original image undergoes a certain transformation. Transform domain coefficients from different sources are fused through a certain rule, and inverse of the transform outputs the fused image. Typical examples of transform domain methods include using Discrete Wavelet Transform (DWT) [1] and Non-Subsampled Contourlet Transform (NSCT) [2]. Transform domain methods are usually simple and fast, but often suffer from loss of original information, with pixels affected by multiple transform coefficients.

In spatial domain approach, certain division rule is applied to the original images. A focus measurement is computed for each of the divisions. Divisions from source images are combined into the fused image based on the focus measure. Typical examples of spatial domain method include energy of image gradients [14], and quad-tree based focus measure [3]. Spatial domain methods can avoid loss of original information by using a maximum selection fusion, and avoid blocking problem by further division of boundary blocks. However, such methods can cause wrong decisions in large flat regions or sharp edges around the decision boundary.

In machine learning based approaches, neural network models are used to determine the activity measurement. Models are trained from large set of labeled image patches for focus measure output. Unseen images are divided into patches and processed by the model for its focus measure, which then guides the fusion of the patch pair. A typcal example of machine learning method is CNN-based fusion technique [4]. As data-driven models, machine learning methods are robust against complex real world scenario. However, labeled training data is hard to obtain, and out-of-focus blur is usually artificially simulated. High Computational power is also needed.

We propose a joint decision algorithm, modified from original algorithm proposed in work [5], to utilize different focus measure methods in the final fusion decision. The original algorithm and our modification are explained in detail in section II. Our modified algorithm is experimented and analyzed on different objective metrics and compared with current state-of-the-art algorithms and the original algorithm in work [5] in section III. We also test the robustness of the proposed algorithm against different levels of artificial noise. Finally, the proposed algorithm is concluded in section IV.

II. Proposed Algorithm Detail

In work [5], the authors have proposed a pixel-based fusion algorithm based on two different activity measures. The input images are divided into high frequency and low frequency parts. Fig. 2b shows the low-frequency part of Fig. 2a, the near-focus source image of a pair; while Fig. 2c shows the high-frequency part, enhanced by multiplying 10. The high-frequency sub-image pixels are transformed using independent component analysis (ICA) bases with a 7x7 window, and the sum of transform domain coefficients is used as the pixel's activity measure.

The low-frequency sub-image pixels, on the other hand, used sum of modified Laplacian (SML) with 3x3 window as activity measurement in work [5], with modified Laplacian (ML) defined as

$$\begin{aligned} ML(x,y) = &|2I(x,y) - I(x,y-1) - I(x,y+1)| \\ &+ |2I(x,y) - I(x-1,y) - I(x+1,y)| \end{aligned} \quad (1)$$

where $I(x,y)$ is the pixel value at location with index x and y; and SML defined as the sum of MLs in a given window:

$$SML(x,y) = \sum_{\substack{a\in[x-1,x+1] \\ b\in[y-1,y+1]}} ML(a,b) \quad (2)$$

As a well-performing measure, we kept this algorithm for low-frequency processing in our proposal.

To combine results from different focus measures, the outcome from different source images are normalized by softmax function, which generates a confidence measure regarding the

pixel's clarity. Fig. 2d shows the low-frequency processing outcome from Fig. 2b, with brighter pixel indicating greater certainty that respective pixel in Fig. 2a is in-focus. Gray pixels means low certainty in the choice between two input images.

The softmax normalized activity measures are multiplied together for the final focus measure. Maximum-selection rule is applied, and the pixel with the highest joint focus measure is selected to present in fused image.

A smoothing algorithm based on minimum graph-cut from work [6] is used to correct wrong decisions in flat area. While node energy is represented by negative log of the joint activity measurement, the edges between two pixels are defined as (3):

$$E_s(m,n) = dist^{-1}(m,n) \cdot \frac{\sqrt{(J_{1,m}-J_{2,n})^2+(J_{1,n}-J_{2,m})^2}}{\sqrt{(J_{1,m}-J_{1,n})^2+(J_{2,m}-J_{2,n})^2}} \quad (3)$$

where m, n indicate two different pixel locations, and $J_{a,b}$ denotes the norm of pixel from source a at location b. The smoothing algorithm will find a division that minimizes sum of edge energybeing cut and the node energy of the decision map. The original method proposed a fully-connected graph [5], with each pixel connected to each other. For memory and processing time consideration, we used a neighbourhood of 21x21 size, and considered edge energy between farther pixels as insignificant. Fig. 2g shows effect of the refinement process from Fig. 2f. The refinement removed much of the noisy blocks in the decision map.

We advocate the framework presented in work [5] as a good method to combine results from different focus measure algorithms and utilizing their respective strength. In fact, softmax normalization of focus measurements from different scale can be applied to most existing and novel image fusion algorithms, as fusion rules are predominantly maximum-selection or weighted average. As such, we implemented this framework with different choices of high frequency solutions and kept the low frequency solution unchanged. We looked both into Laplacian Pyramid, a transform domain method, and image variance, a spatial domain method. Their implementation and integration are discussed in section II-A and II-B respectively.

To incorporate the need for noise-free input for some activity measures, we also added a denoising module before dividing the image according to frequency. A novel image denoising algorithm Swin-Conv-UNet [8] is used. Our implementation is detailed in section II-C.

The information distribution across different frequency varies from image to image while condense in low frequencies. For some input images, little to no pixel value exist in high-frequency sub-image with fixed cut, and high-frequency part analysis comes back inconclusive. To provide enough information for high-frequency algorithm to process, a binary search elaborated in section II-D is implemented to cut at the frequency where enough data is preserved in high-frequency sub-image.

The process flow diagram is shown in Fig. 1. A fused example is shown in Fig. 2h.

A. Laplacian Pyramids

Laplacian pyramid is a type of transform domain method - transform the input image into an alternative domain, fuse in that domain, then transform back. Laplacian pyramid is constructed by taking the difference between two levels of a Gaussian pyramid. Difference operation computes the band-pass filtered versions of the image, capturing different levels of spatial frequencies or details. This is similar to band-pass filtering the image.

These difference images (levels of the Laplacian Pyramid) are conceptually similar to applying the Laplacian operator on the original image, which emphasizes the high-frequency components or spatial variations. While the Laplacian Pyramid does not explicitly use the Laplacian operator, the name reflects the shared motivation of capturing the spatial variations or high-frequency details in the image data. The focus measure of the Laplacian pyramid is the amplitude of the node in the layer of the pyramid. When a segment of the image is out of focus, high frequency energy is lost first. Therefore, the amplitude of the node directly corresponds to how high-frequency or high-activity that pixel is.

Two significant issues with Laplacian pyramids are addressed with the proposal. First, as a bandpass fliter, it should not pass noise, but it does not deal with high frequency noise well. Gaussian and salt & pepper noise usually pass through. This limitation is resolved with the Swin-Conv-Unet Denoising Module in section II-C. Secondly, as a shift-variant transform, a slight misalignment between the two source images greatly affects its in-focus region identification. In this report, we assumed the no image mis-registration.

B. Image Variance

Variance measures how much change there is within a group. If all the elements in the group are the same value, there is no variance. If the values are very different from the group's average, the variance is high. When applied to images, variance is a good measure of activity within the image. A purely red image would have a very low variance, while an image of a person's face has higher variance. Variance calculations are useful in image processing to determine the focal areas of an image. As shown in Fig. 3, the input image is split into partitions and then the variance of each partition is calculated as a focus measure. A high variance for a partition means a high level of focus for that area. Variance for partition i,j with width M and height N is calculated as follows [7]:

$$Variance(i,j) = \frac{1}{M*N}\sum_x\sum_y (f(x,y)-u)^2 \quad (4)$$

where u:

$$u = \frac{1}{M*N}\sum_x\sum_y f(x,y) \quad (5)$$

This focus measure is then compared to the image with a different focal length. The two calculated partition variance from different images is passed into softmax function for a possibility measurement P_h regarding the partition being

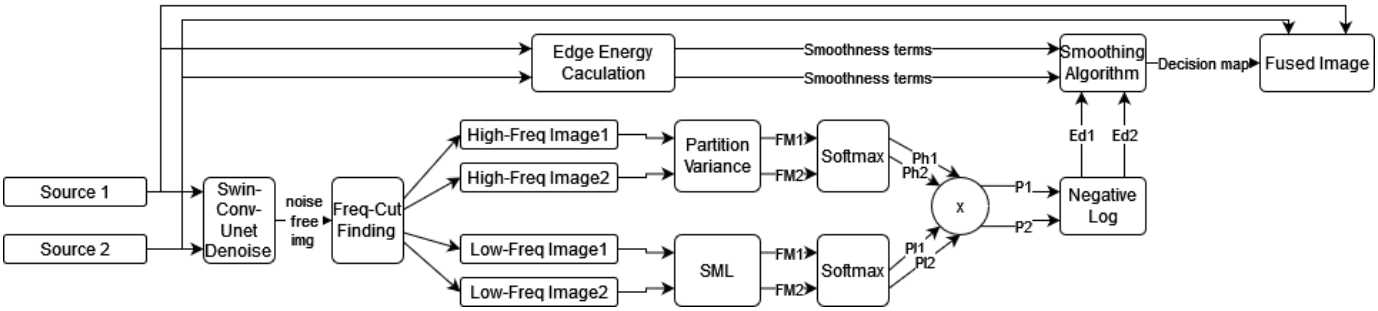
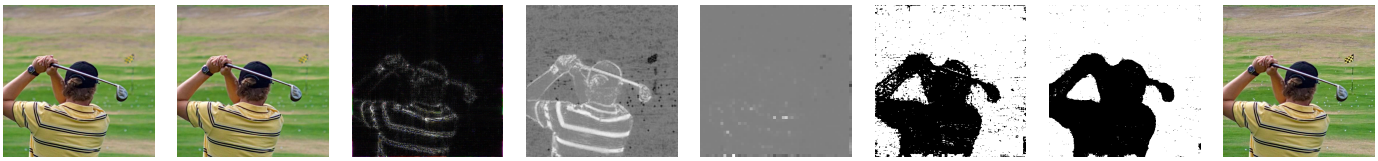


Fig. 1: Proposed algorithm workflow



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 2: example image processed under proposed algorithm

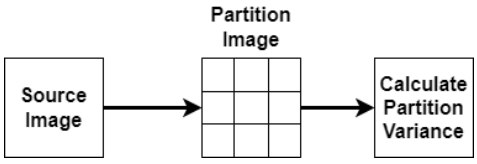


Fig. 3: Image Partitions and Variance Calculations

present in fused image. The partition variance processing result for the high frequency component shown in Fig. 2c is visualized in Fig. 2e. We can see that whiter blocks, indicating higher centainty of choosing the block to present in fused image, form the shape of the man in the original image Fig. 2a, which is the in-focus region for the image.

A partition size is chosen so that each partition contains enough information to obtain a useful variance metric while also not having the partition too large to cover unique sections of the image. A partition size of 10x10 pixels was chosen due to such partitions results in more accurate image fusion. Images that cannot be partitioned into equal 10x10 sections can be padded with zeros since they will have the least impact on the variance.

C. Swin-Conv-UNet Denoising Module

The accuracy and effectiveness of the processing activities might be considerably impacted by the image noise. Noise in the input image can degrade the fusion process and reduce the quality of the final image. Denoising can reduce noise and undesired artifacts, which improves the signal-to-noise ratio and enhances the quality of the input images for fusion.

The Swin-Conv-UNet algorithm is an effective approach to removing noise in real-world scenarios. High-frequency images typically contain more detailed and textured features, which can be affected more by noise and may require more complex denoising techniques to preserve. The Swin Transformer-based backbone network used in the algorithm is designed to handle large and complex images efficiently, thanks to its hierarchical and multi-scale feature extraction capabilities. Moreover, the use of data synthesis techniques for training can help to simulate various noise and image conditions and improve the algorithm's robustness and adaptability.

The algorithm works in the following stages. The first stage involves generating synthetic data for training purpose. The original image is synthesized with various noise and image conditions to form a large training dataset to improve the algorithm's robustness and accuracy. The second stage involves normalizing and resizing the image. The normalized image is further fed to the Swin Transformer-based backbone network which is a convolutional neural network that extracts hierarchical and multi-scale features from the image. The network is trained on a large dataset of synthetically generated noisy images from our first stage. The extracted features are then fused using the U-NET architecture that combines low-level and high-level features to enhance image representation and reduce the effect of noise. The fused image is then post-processed by applying a pixel-wise nonlinear function to the output of the network to reduce residual noise and improve the overall quality of the denoised image. For the scope of this project, a pre-trained model is used.

D. Adaptive High-Low Frequency Cut

To keep high frequency image with enough information for effective activity measure and fusion decision, a binary search algorithm is implemented find the appropriate cut such that high-frequency portion include large enough total pixel values. After the fourier transformed image is centered for a frequency range $[-\pi, \pi]$, the cut is initialized at midpoint, where low frequency include frequency components under $\pm\pi/2$ and

high frequency means frequency components above $\pm\pi/2$. The high frequency portion is transformed back to spatial domain, and the total pixel value is evaluated and compared with original image. If the cut does not meet the certain percentage range of original image, depending on being too large or too small, the search algorithm will try the midpoint of upper or lower half respectively. The search stops when a cut that meet the requirement is found. For our algorithm, the percentage range is set to $[0.5\%, 1.5\%]$.

III. Metrics and Experiments

A. Evaluation Metrics

In work [9], Liu et al. generalized image fusion metrics into 4 categories, based on mutual information, gradient preservation, structural similarity, and human perception process. In this article, we chose one metrics from each of the four categories to evaluate our proposed algorithm. Mutual Information metric MI proposed by Hossny et al. [10] evaluates the joint entropy against individual entropy of images for information retained. Gradient-based metric Q_G proposed in work [11] estimates how edges are preserved. Structural similarity metric Q_Y from work [12] compares statistical property of pixels in their neighbourhood between fused and input images. Human perception inspired metric Q_{CB} by work [13] measures the contrast transferred. For all of the metrics, higher value indicates better fusion.

B. Comparison with Other Methods

A comparison between our proposed method and current state-of-the-art methods are shown in Tab. I, with the number in bracket indicating our method's rank. For the two proposed high frequency solutions, we equally split the dataset to implement each of them. We compare our method against two transform domain methods: DWT-based [1] and NSCT-based [2]; two spatial domain methods: MWGF-based [14] and quadtree-based [3]; and one neural network method CNN [4]. The methods are compared based on the metrics listed in section III-A. We used Lytro Image Fusion Dataset [15] for this comparison. The implementation of the metric score algorithm and metric scores for state-of-the-art fusion algorithms are provided by Liu et al. in work [9].

TABLE I: Mean fusion metrics against Lytro Dataset

Methods	MI	Q_G	Q_Y	Q_{CB}
DWT	0.9019	0.7357	0.9545	0.7362
NSCT	0.9523	0.7454	0.9618	0.7491
MWGF	1.1278	0.7475	0.9873	0.7978
Quadtree	1.1864	0.7609	0.9886	0.8095
CNN	1.1512	0.7615	0.9875	0.8084
mf-CRF[a]	1.1188[b]	0.7641	0.9896	0.8098
Proposed Alg.	1.1836(2)	0.7595(4)	0.9867(5)	0.8006(4)

[a] Mean metrics of this method come from [5]
[b] Rescaled from original report 8.9506, unreliable conversion

From the metrics, our method is competitive to recent published methods and our main referenced publication [5], even though older techniques such as Laplacian Pyramids is used. While our method retains the advantage in mutual information with a maximum selection fusion rule, other metrics falls slightly behind recent methods. This could attribute to our simplification in the smoothing algorithm for resource considerations. Work [6] proved convergence only if all the edges are connected, but our simplification to a windowed neighbourhood made it not converge. Early stopping causes less refined decision map, and thus reduced gradient preservation and structral similarity.

C. Performance against Noise

For testing our model against different levels of noise, we used 5 image pairs from Lytro Image Fusion Dataset [15] and added different types of artificial noise in the following sequence: i) Poisson noise; ii) multiplicative speckle noise; iii) Gaussian additive white noise (AWGN); iv) salt & pepper noise. The noises are adjusted through a parameter affecting the severity of the interference. At level 5, AWGN has a variance of 0.01; speckle noise has a variance of 0.05; and salt & pepper noise affect 5% of the image. We consider the noise-free fusion result from our algorithm as the ground truth fusion, and calculate PSNR and MSE of the fusion result from noisy image against them. These metrics shows how the fusion quality decay as greater noise is added.

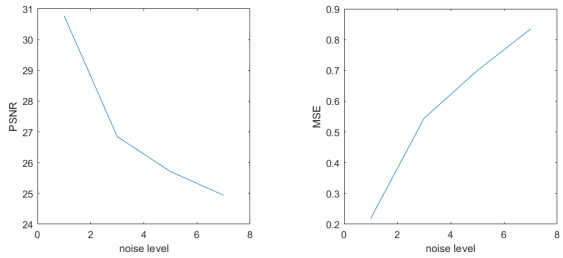


Fig. 4: Proposed algorithm against different noise levels

From the plot, our method can keep PSNR above 20 even under severe noise, with MSE staying below 1. We conclude that the algorithm is effective in noisy environment. However, it must be noted that the noise added are artificial.

IV. Conclusion

In this paper, a novel joint image fusion algorithm is proposed which leverages the advantages of different existing image fusion techniques including adaptive high-low frequency separation, Laplacian pyramids, image variance, and signal denoising. Experiment results show competitive accuracy and quality against recent published fusion methods. The fusion algorithm is tolerant to heavy noise. The proposed joint decision framework is highly flexible to be utilized in combining most existing fusion methods. The implementation used in this report has compromised greatly with limited computation resources, and great improvement on processing time and decision accuracy is achievable by utilizing GPU acceleration and large memory. The novel modified image fusion technique should be implemented in fusion problems where the quality is prioritized over speed.

REFERENCES

[1] H. Li, B. S. Manjunath, and S. K. Mitra, “Multi-sensor image fusion using the wavelet transform,” in Proc. 1st Int. Conf. Image Process., vol. 1, Nov. 1994, pp. 51–55.

[2] Q. Zhang and B.-L. Guo, “Multifocus image fusion using the non-subsampled contourlet transform,” Signal Process., vol. 89, no. 7, pp. 1334–1346, 2009.

[3] X. Bai, Y. Zhang, F. Zhou, and B. Xue, “Quadtree-based multi-focus image fusion using a weighted focus-measure,” Inf. Fusion, vol. 22, pp. 105–118, Mar. 2015.

[4] Y. Liu, X. Chen, H. Peng, and Z. F. Wang, “Multi-focus image fusion with a deep convolutional neural network,” Inf. Fusion, vol. 36, pp. 191–207, Jul. 2017.

[5] O. Bouzos, I. Andreadis and N. Mitianoudis, ”Conditional Random Field Model for Robust Multi-Focus Image Fusion,” in IEEE Transactions on Image Processing, vol. 28, no. 11, pp. 5636-5648, Nov. 2019, doi: 10.1109/TIP.2019.2922097.

[6] Y. Boykov, O. Veksler, and R. Zabih, “Fast approximate energy minimization via graph cuts,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, no. 11, pp. 1222–1239, Nov. 2001.

[7] Wei Huang and Zhongliang Jing, ”Evaluation of focus measures in multi-focus image fusion” in Pattern Recognition Letters, vol. 28, no. 4, pp. 493-500, 2007

[8] Kai Zhang, Yawei Li, Jingyun Liang, Jiezhang Cao, Yulun Zhang, Hao Tang, Radu Timofte, Luc Van Gool, ”Practical Blind Denoising via Swin-Conv-UNet and Data Synthesis” arXiv:2203.13278v2

[9] Z. Liu, E. Blasch, Z. Xue, J. Zhao, R. Laganiere and W. Wu, ”Objective Assessment of Multiresolution Image Fusion Algorithms for Context Enhancement in Night Vision: A Comparative Study,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 1, pp. 94-109, Jan. 2012, doi: 10.1109/TPAMI.2011.109.

[10] M. Hossny, S. Nahavandi, and D. Creighton, “Comments on ‘Information measure for performance of image fusion,”’ Electron. Lett., vol. 44, no. 18, pp. 1066–1067, Aug. 2008.

[11] C. S. Xydeas and V. Petrovic, “Objective image fusion performance measure,” Electron. Lett., vol. 36, no. 4, pp. 308–309, Feb. 2000.

[12] C. Yang, J.-Q. Zhang, X.-R. Wang, and X. Liu, “A novel similarity based quality metric for image fusion,” Inf. Fusion, vol. 9, no. 2, pp. 156–160, 2008.

[13] Y. Chen and R. S. Blum, “A new automated quality assessment algorithm for image fusion,” Image Vis. Comput., vol. 27, no. 10, pp. 1421–1432, Sep. 2009.

[14] Z. Zhou, S. Li, and B. Wang, “Multi-scale weighted gradient-based fusion for multi-focus images,” Inf. Fusion, vol. 20, pp. 60–72, Nov. 2014.

[15] M. Nejati, S. Samavi, S. Shirani, ”Multi-focus Image Fusion Using Dictionary-Based Sparse Representation”, Information Fusion, vol. 25, Sept. 2015, pp. 72-84.