Computers, Materials & Continua
DOI:10.32604/cmc.2022.021125
Article

Tech Science Press

Fusion of Infrared and Visible Images Using Fuzzy Based Siamese Convolutional Network

Kanika Bhalla[1], Deepika Koundal[2,*], Surbhi Bhatia[3], Mohammad Khalid Imam Rahmani[4] and Muhammad Tahir[4]

[1]Department of Electrical Engineering, National Taipei University of Technology, Taipei, 10608, Taiwan
[2]Department Virtualization, School of Computer Science, University of Petroleum & Energy Studies, Dehradun, India
[3]College of Computer Science and Information technology, King Faisal University, 36362, Saudi Arabia
[4]College of Computing and Informatics, Saudi Electronic University, Riyadh, 11673, Saudi Arabia
[*]Corresponding Author: Deepika Koundal. Email: koundal@gmail.com
Received: 24 June 2021; Accepted: 03 August 2021

Abstract: Traditional techniques based on image fusion are arduous in integrating complementary or heterogeneous infrared (IR)/visible (VS) images. Dissimilarities in various kind of features in these images are vital to preserve in the single fused image. Hence, simultaneous preservation of both the aspects at the same time is a challenging task. However, most of the existing methods utilize the manual extraction of features; and manual complicated designing of fusion rules resulted in a blurry artifact in the fused image. Therefore, this study has proposed a hybrid algorithm for the integration of multi-features among two heterogeneous images. Firstly, fuzzification of two IR/VS images has been done by feeding it to the fuzzy sets to remove the uncertainty present in the background and object of interest of the image. Secondly, images have been learned by two parallel branches of the siamese convolutional neural network (CNN) to extract prominent features from the images as well as high-frequency information to produce focus maps containing source image information. Finally, the obtained focused maps which contained the detailed integrated information are directly mapped with the source image via pixel-wise strategy to result in fused image. Different parameters have been used to evaluate the performance of the proposed image fusion by achieving 1.008 for mutual information (MI), 0.841 for entropy (E_G), 0.655 for edge information (EI), 0.652 for human perception (HP), and 0.980 for image structural similarity (ISS). Experimental results have shown that the proposed technique has attained the best qualitative and quantitative results using 78 publically available images in comparison to the existing discrete cosine transform (DCT), anisotropic diffusion & karhunen-loeve (ADKL), guided filter (GF), random walk (RW), principal component analysis (PCA), and convolutional neural network (CNN) methods.

Keywords: Convolutional neural network; fuzzy sets; infrared and visible; image fusion; deep learning

1 Introduction

The infrared sensors or multi-sensors are used to capture the infrared and visible images. As different objects like the environment, people, and animals emit thermal or infrared radiations which are further used for the detection of target and parametric inversion. These images have a lesser effect and insensitive to the illumination variations and disguise. Thus, it overcomes the hurdles while detecting the targets by working day and night [1]. But the most important visible feature such as texture information get lost due to the small spatial resolution of the infrared images, as a result objects contain insufficient details. This is due to the temperature-based nature of the object. The objects that are warmer and colder than the background is easier to detect. On the contrary, visible images deal with the spectral resolution and sensitivity to the effect of changing brightness or illumination in the scene. It illustrates the perceptual scenes for the human eyes and human vision system (HVS) [2]. Sharpened high spatial resolution VS images depicts the important information on the geometric details of the objects, and thus helps for overall recognition [3]. But mostly, a target cannot be easily identified due to the changing environmental and poor lighting conditions like objects covered in smoke, disguises, night time, and disordered background. Sometimes background and targets are similar due to which obtained information is insufficient. Hence, IR/VS images offer integrative advantages.

Therefore, there is a need for an automatic fusion method that can fuse the two complementary images into a single image, i.e., integration of thermal radiations of the IR and texture appearance of the VS images to produce an enhanced vision of an image [4,5]. Furthermore, the main aim is to obtain the fused image with abundant VS image details and chief thermal targets from the IR images. Hence, the goal of the IR/VS image fusion is to preserve the useful features of IR and VS images.

In recent years, more attention has been paid towards the field of IR and VS image fusion. Many researchers presented a lot of IR/VS image fusion approaches which are roughly classified into various categories as multi-scale decomposition (MST), principal component analysis (PCA), sparse representation (SR), fuzzy sets (FS), and deep learning (DL). In consideration to this problem, the main motivation behind this work was to extend the research in the direction of the examination of the fused image to be helpful in the object tracking, object detection, biometric recognition, and RGB-infrared fusion tracking. Therefore, goal is to propose a reliable automatic anti-noise infrared/visible image fusion technique for generating a fused image that has the largest degree of visual representation of environmental scenes to be used in.

Major contributions of this study are: (1) The unique integration of fuzzification and siamese CNN based infrared/visible fusion technique for the integration of complementary infrared/visible images has been put forward. (2) Fuzzification has been done using the fuzzy sets to model various uncertainties efficiently for problems like ambiguousness, vagueness, unclearness, and distortion present in the image by the determination of the membership grade of the background environment as well as target detection. Whereas, feature classification has been done by the CNN model with the extraction of the low level as well as high level infrared/visible features. Furthermore, fusion rules are also automatically generated to fuse the obtained features. (3) The proposed technique is more reliable and robust as compared to the classical infrared/visible technique due to its advantage of making it less laborious. (4) A publically accessible dataset consisted of 78 infrared/visible images has been used for the experiments. (5) The qualitative as well as quantitative evaluation has been done using six classical infrared/visible techniques such as discrete cosine transform (DCT), anisotropic diffusion & karhunen-loeve (ADKL), guided filter (GF), random walk (RW), principal component analysis (PCA), and convolutional neural

network (CNN) methods by using five metrics, i.e., mutual information (MI), entropy (E_G), edge information (EI), human perception (HP), and image structural similarity (ISS). Higher results are given by the proposed technique proves its effectiveness concerning other pre-existing techniques. This study deals with the problem of pixel level multi-sensor image fusion.

The key motivation of this research is to combine the advantages of the spatial domain (CNN) and fuzzy based method to achieve the accurate extraction of IR targets while maintaining the background features of VS images which is not easy to attain as there occurs various challenges during this process. Efficacious evaluation of the quality of pixels has been done with the extraction of target features and background features in order to integrate them for the generation of clear focused fused image. Additionally, it is a laborious task. Then, investigation of the determination of the pixels belongingness is an issue of relevance. Furthermore, from the literature, it has been analyzed that FS represented the uncertain features. Therefore, indeterminacies, noise, and imprecision present in the images can be considered as a problem of fuzzy image processing. Subsequently, due to the powerful ability of the CNN for automatic data extraction, this research work generated the data-driven decision maps with the utilization of CNN. Hence, as per the literature, no attempt has been made to integrate the FS with CNN for IR/VS image fusion. Therefore, in this research work, an attempt has been made to propose a novel fuzzy CNN based IR/VS image fusion method for the fusion of images. The key contributions of this study are outlined as follows.

- It helped to integrate different modality images to produce a clear more informative fused image.
- It also improved the infrared image recognition quality of the modern imaging system.
- Subjective and objective experimental analysis have been performed.

The remaining structure of this study is presented as follows: Section 2 briefly describes the background and related approaches for infrared/visible image fusion. In Section 3 detailed description of the proposed technique methodology is given. Section 4 presents the dataset, evaluation metrics, and validates the experimental results by doing an extensive comparison with existing techniques. In Section 5, concluding remarks and future works discussion is drawn.

2 Related Works

In the past, numerous techniques for infrared/visible fusion had been developed like pyramid decomposition [6], and DCT [7]. But, they were not suitable methods as they produced oversampling, high redundancy, and so many other problems. Whereas, histogram-based methods [8,9] produced unsatisfactory results due to their inability to amplify gray levels of the images as well as background distortion. Hence, they produced the low-quality fused images. Bavirisetti et al. [10] introduced the edge preserving ADKL transform technique. Although, good results were obtained but still the qualitative as well as quantitative results needs to be improved and along with this it was a labor-intensive method. Further, Liu et al. [11] presented the convolutional sparse representation method whose main drawback was that only the last layer was used for the extraction of features which resulted in the loss of the most useful information. Hence, it was a crude method. Liu et al. [12] developed a variation model which was based on saliency preservation. Only seven image sets were used, which were the main limitation of this study. Many non-subsampled contourlet transform (NSCT) approaches [13,14] were developed. However, these methods gave satisfactory fused images but there were many drawbacks like the process was cumbersome and tedious. The decomposition of the image and reconstruction of the fused image was computationally intensive and was not a feasible method to be used in real time applications.

Yang et al. [15] developed the guided filter (GF) technique for the measurement of visual features of the image. Although a better-quality of fused image was obtained but still subjective and quantitative results need to be enhanced. Only five sets of infrared/visible images with three evaluation metrics were used for the validation purpose. Ma et al. [16] developed a boosted RW method for the effective estimation of the two-scale focus maps. The quality of the fused image needed to be improved. Afterwards, Shahdoosti et al. [17] introduced a hybrid technique with an integration of PCA and spatial PCA techniques with the usage of an optimal filter. Subsequently, obtained the synthesized results similar to the corresponding multisensors observed at the high-resolution level. Liu et al. [18] developed the DL framework for the integration of multi-focus images which was computationally intensive. They have exhibited their applications on other types of modalities such as infrared/visible image fusion. Many other DL based techniques were introduced for the fusion of different modality images. Li et al. [19] presented a fusion framework based on DenseNet. Four convolution layers were included in the encoder block. Shallow features were extracted by one of the convolutional layers. Another three layers constituted the Dense block were used to obtain both shallow and deep image features. Then, Li et al. [20] fused the visible and infrared images using VGG network. In this approach, middle layer information were utilized but the information loss during the integration of features limited the model's performance. Ma et al. [21] propounded an image fusion method based on generative adversarial network (GAN). Whereas, adversarial network was adopted to extract more visible details of the images. Zhang et al. [22] designed a transform domain based convolutional neural network approach constituting both feature extraction and reconstruction blocks. In this architecture, 2 CNN layers utilized to obtain features of an image for fusion, and then reconstructions of image features were done to generate fused images. Xu et al. [23] developed the U2Fusion architecture for the fusion of images. This method was based on DenseNet [24] where vital information were retained by the designed information measurement. Zhao et al. [25] attained the fused images by the designing of self-supervised feature adaption architecture. Moreover, fuzzy set-based approaches [26–28] were also used due to their very strong mathematical operations to deal with the fuzzy concepts even whose quantitative illustration was not possible. Thus, on the basis of the literature study, the above limitations motivated us to hybridize the advantages of fuzzy sets with deep learning concepts. The presented work focused on the development of an automatic effective infrared/visible image fusion technique for enhancing the vision of the fused image. With this incorporation, this method preserved vital information.

3 Proposed Methodology

In order to handle the former problems, hybridization of the fuzzy set and Siamese CNN has been employed to fuse the infrared/visible images. The proposed technique is presented as follows.

3.1 Fuzzification

Zadeh et al. [29] introduced the concept of a fuzzy set which is a very useful mathematical expression to handle an object with some kinds of imprecision and uncertainties like distortion, vague boundaries, ambiguity, blurriness, uneven brightness, and poor illumination [30]. When the infrared/visible images are captured by sensors, there occurs an ambiguity in image pixels. Their belongingness to the target or background is considered to be a typical problem. Therefore, this problem has been solved by the use of fuzzy sets that further helps to solve the existence of intermediate values by the assignment of a degree of truth ranges from 0 to 1 typically deals with an uncertain problem.

For the processing of an image, input images L and M was converted from pixel domain to fuzzy domain. Eq. (1) illustrated the image representation. Let's assume an image L as for illustration. $L(i,j)$ implies its pixel values whose mapping has to be done into its fuzzy characteristic domain. It has expressed as shown below by Eq. (1).

$$L(i,j) = \sum_{i=1}^{p} \sum_{i=}^{q} [\mu L(i,j) / L(i,j)] \tag{1}$$

where $\mu L(i,j)$ is a membership degree, whose values range from 0 to 1 i.e., $\mu L(i,j)$: $L \to [0,1]$, L is an element of the universal set. Each pixel is represented by $\mu L(i,j) / L(i,j)$. Therefore, mapping of the original pixel value (0 to 255) is mapped to (0 to 1) i.e., fuzzy plane.

The membership grade describes the element's degree of belongingness to a FS. Here, 1 indicates the elements with complete belongingness to a FS, whereas 0 implies it's belonginess to the fuzzy set. Summation of all the membership functions of the element 'L' is 1 as represented below.

$$\sum_{i=1}^{p} \mu_L(i,j) = 1 \tag{2}$$

where, p represents the number of FS where L belongs.

As input grayscale image includes darker, brighter, and gray level pixels whose value ranges from 0 to 255. Therefore, image mapping has been done from pixel scale to fuzzy domain by assigning triangular membership function.

Now, image having pixel values between 0 to 255 was converted to 0 to 1 indicating the pixel fuzziness.

$$\mu_{(L(i,j))} = \begin{cases} 0, & \text{if } \mathrm{F} < L(i,j) \le h \\ (L(i,j) - \mathrm{F}) / (g - \mathrm{F}), & \text{if } \mathrm{F} < L(i,j) \text{ and } L(i,j) \le g \\ h - L(i,j) / h - g, & \text{if } g < L(i,j) \le h \\ 1, & \text{if } L(i,j) > h \end{cases} \tag{3}$$

where F, h, and g implies the minimum, average, and maximum pixel intensity value, respectively. $L(i,j)$ indicates the input image pixel value.

The triangular membership function of the image has been applied whose mathematical representation is shown in Eq. (3). Now, the image having pixel values between 0 to 255 is converted to 0 to 1 indicating the pixel fuzziness. The membership grade describes the element's degree of belongingness to a fuzzy set. Here, 1 indicates the elements with complete belongingness to a fuzzy set, whereas 0 implies that it does not belong to the fuzzy set. The calculation of membership value i.e., the process of fuzzification is given by Eq. (3).

So, by using the above equation, pixels having minimum intensity value are assigned 0 whilst pixels having maximum value are assigned 1, and the uncertainty, as well as ambiguity are removed. Hence, the uncertainty was removed without diminishing the image quality.

3.2 Siamese CNN

The proposed Siamese CNN or convNet model designed for the fusion of IR/VS images is described here. It is designed to automatically learn mid and high level abstractions of the data

presented in the two heterogeneous images. By the use of the Siamese network, the same weights were shared between two different branches. One branch was used to handle the infrared image and the other was to process the visible image. Each branch has step-wise stages of CNN such as convolution layer, max pooling, flattening, and full connection, i.e., fully connected layer.

These layers generate the feature maps parallel to each level of abstraction of features from an image [31]. The CNN framework configuration for infrared/visible image fusion is used as a stack of varied convolutional layers consisting of 3 convolutional layers, one max pooling, two FC layers, and one output softmax layer. Therefore, the above-discussed features have been captured by using three convolutional filters with a feature detector size of 3×3 pixels. ReLU is slid over the whole input volume. ConvNet has been used during the implementation of the proposed technique which has each layer of the convolution composed of (a) 3×3 convolutions (b) Batch Normalization (BN′) (c) ReLU function and (d) max pooling.

Then, features extracted from the previous CNN layers are concatenated by the fully connected (FC) layer. Subsequently, pooled feature maps are obtained by the flattening of the pooling layers. The last layer consists of the output neuron which assigns a probability to the image. CNN gives scalar output whose value ranges from 0 to 1.

3.3 Fusion Scheme

The proposed technique for the fusion scheme consisted of five steps: fuzzification, focus detection by the feature map generations, segmentation, unwanted region removal, and infrared/visible image fusion. This attempt has been made to generate a fused image consisting of all its useful features as illustrated by the schematic block diagram of the proposed technique for infrared/visible image fusion in Fig. 1.

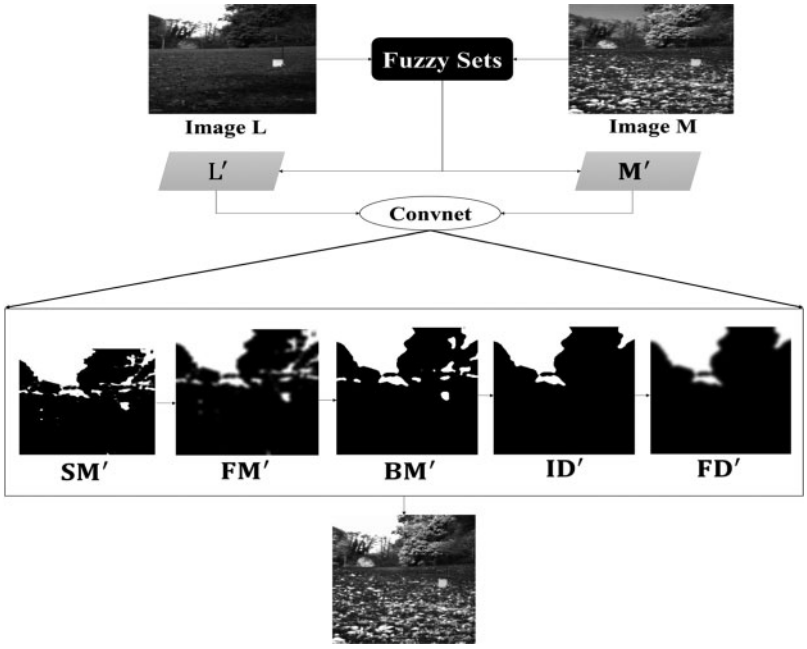

Figure 1: Schematic block diagram of the proposed infrared/visible image fusion

Firstly, L and M, the infrared/visible images, respectively are fed to the fuzzy set. Then, the fuzzification has been done by doing the processing of the information presented in the image, followed by L' and M', then, the output fuzzified images are passed to the pre-trained Siamese CNN model. Additionally, binary classification of the infrared/visible images has been done by pretrained Siamese CNN. During this, various distinct extracted image pixels are transferred to the next convolutional layer until the entire classification is done.

For the first three convolutional layers, the fixed stride of 1 has been used. Max pooling has been applied for the localization of the parts of the images using a window size of 2×2. This has a stride of two, further helped in choosing the larger pixel value from each part of an image. Therefore, every time the CNN layers stack is followed by the ReLU to keep the constant network volume. In summary, the first three convolutional layers are represented as Con1, Con2, and Con3. Con1 has generated the 64 FM' after applying the 64 filters, Con2 has produced the 128 FM' by the use of 128 filters. Due to the self-learning feature of the CNN, these filters are automatically applied. After that, 256 FM' are obtained which then passed to the FC layer for further combining with 256 dimensional vectors to produce two dimensional output vector. Lastly, probability distribution among the two classes has been obtained using the SoftMax function. These are followed by the generation of feature maps. The main task of the CNN is to do automatic feature extraction from the given input images.

Thus, during the fusion, the network which has been trained using the patch size of 16×16 is fed with 2 fuzzified source images to generate a score map SM'. In detail, first Con 1 has extracted only low i.e., dark feature maps having high frequency information from the images. Therefore, to capture the spatial detail of the image, Con 2 has been added. It has produced the feature maps covering varied gradient orientations. Hence, the third convolutional layer has integrated the gradient information and produced the output feature map i.e., score map. Here, SM' illustrated the focused information describing the image focus ability of a set of patches having 16×16 size in the source image whose value ranges from 0 to 1. A more detailed focused patch has been obtained when its value is near 0 (black) and 1 (white). SM' size is given in Eq. (4).

$$SM' = \left(\left\lceil \frac{ht}{2} \right\rceil - conv\_patchsize + 1 \right) \times \left(\left\lceil \frac{wt}{2} \right\rceil - conv\_patchsize + 1 \right) \tag{4}$$

If $0 < SM' < 1$, it implied the focused parts, ht and wt implies the height and width of the image respectively. Here, there is a reduction of the size of $(SM)'$ to half due to the presence of overlapping pixels. Therefore, ht, wt, and $conv\_patchsize$ is also reduced to half and conv_patchsize was reduced to 8×8.

Moreover, SM' consisted of an overlapped pixel. Hence, an averaging method was utilized to produce a focus map of the same size to the source image. Now, focused information is correctly detected where the black or white region represents the more abundant detailed image information. However, the plain regions (gray) constitute a value of 0.6. To generate accurate focused map, threshold factor of 0.6 was chosen empirically to keep the balance between good quality and computationally expensiveness. As this is an optimum value chosen which gives the

best binary segmentation map with the best evaluation metrics results in comparison to the other values.

$$FM' = \begin{cases} 1, White \\ 0, Black \\ 0.6, Gray \end{cases} \tag{5}$$

Now, more detailed information is contained in the focus map of the image which is near to 0 or 1 values. From Fig. 2, it can be observed that the obtained focus map constitutes of correctly classified gray pixels, as shown in the white background.

Further processing of the focus map has been done to preserve the maximum of useful features i.e., only to have focused parts i.e., black or white. For this purpose, the maximum method has employed a 0.6 threshold value to segment FM′ to get binary map BM'. As for the segmentation purpose, user defined threshold value, i.e., 0.6 has been selected to obtain the good quality BM′ by the following conditions.

$$BM' = \begin{cases} 1, & if \ \ FM'(x,y) > 0.6 \\ 0, & otherwise \end{cases} \tag{6}$$

The obtained binary map contained some misclassified pixels and unwanted small objects or holes as clearly seen in Fig. 1. Therefore, for the removal of some of the misclassified pixels from the FM' small region removal has been done using bwareaopen to generate the initial decision (ID') map which will produce an image free from unwanted objects by using Eq. (7).

$$ID' = bwareaopen\left(BM', area\right) \tag{7}$$

Here, the area threshold value is manually adjusted to 0.03 i.e., the threshold for area is given in Eq. (8).

$$area = 0.03 \times ht \times wt \tag{8}$$

Now the computed ID' further contained undesirable artifacts on the edges. This has been improved by using edge preserving guided filter. Fig. 2 has clearly justified the difference between the resultant output fused image with and without using a guided filter. The fused image obtained by using the average rule on the ID' without using guided filter containing blurriness whereas the image obtained after using a GF is sharper and brighten. Subsequently, final decision map FD' has been calculated with the use of a guided filter.

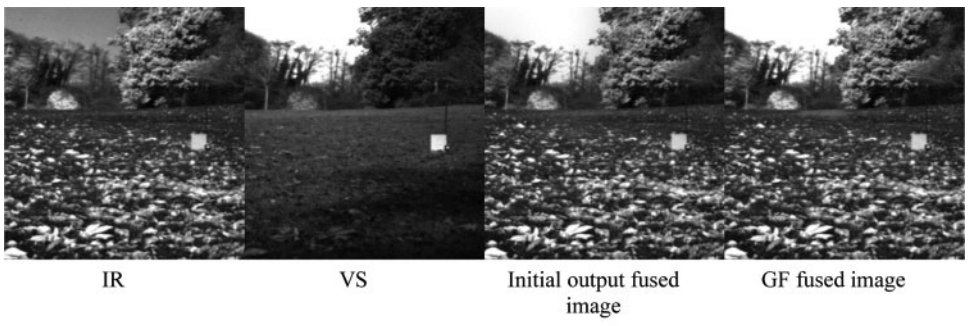

Figure 2: Fused image generated without using GF and with the use of GF

$$FD' = guided\ filter\left(inititial\ fused\ \ image, ID', r, eps\right) \tag{9}$$

where r is set as 5, eps is 0.2. This initial fused image is used as a guidance image for the calculation of FD'.

Lastly, the pixel-wise weighted average method has been used to obtain the resultant single fused image as described in Fig. 1 using Eq. (10).

$$F'(x,y) = FD'(x,y)\,L'(x,y) + 1 - FD'(x,y)\,M'(x,y) \tag{10}$$

where, L' and M' are the given fuzzified images, $FD'(x,y)$ is the final decision map and fused single image is represented by $F'(x,y)$.

The proposed algorithm for infrared/visible image fusion is described in detail in Algorithm 1.

Algorithm 1: Fuzzy CNN based infrared/visible image fusion

Input: Two images, i.e., L and M.
Output: Single fused image F′.
Step 1: Input two images L and M are read.
Step 2: Convert L and M from the gray scale into the fuzzy domain by using fuzzy sets to give L′ and M′, respectively.
Step 3: Compute the SM′ using pretrained Siamese CNN.
Step 4: Generate focus map FM′ using average method illustrated in Eq. (5).
Step 5: Get the binary segmented map BM′ after the calculation of FM′ using Eq. (6).
Step 6: Initial decision map was calculated using Eq. (7).
Step 7: Use GF, for obtaining final decision map (FD′) by setting values of the parameters, r = 5 and $eps = 0.2$ as described in Eq. (9).
Step 8: Output fused image F′ is displayed by Eq. (10).

4 Experimental Evaluations

In this research work, both subjective and objective assessment has been done for the validation of the superiority of the proposed technique. For this purpose, six pre-existing infrared/visible image fusion techniques such as DCT [7], ADKL [10], GF [15], DL [18], RW [16], and PCA [17] have been compared.

4.1 Data Acquisition

In this, IR/VS images are obtained under changing environmental conditions. The publically available datasets are acquired from RoadScene [23], TNO [32], and CVC-14 [33] datasets. The experimental results have been conducted using 78 sets of infrared/visible images. Simulations are conducted in Matlab R2016a, 64-bit using PC with processor Intel® Core™ i5-3470 CPU, 16.0 GB RAM.

The RoadScene dataset consisted of total 221 IR/VS sets of images. Images are of rich road traffic spots. For an instance, pedestrians, roads, and vehicles. These highly representative spots are acquired from naturalistic driving videos. These images have no uniform resolution.

The TNO dataset is common publically used for the IR/VS research. It includes the varied military relevant scenes images that has registered with distinct multi-band cameras with non-uniform resolution.

The CVC-14 dataset included pedestrian scenes that is highly utilized for the manufacturing of autonomous driving technologies. It is composed of two pair of sequence, namely day and night pairs, respectively. Total are 18710 images, among which 8821 is the daytime sequence and 9589 as the nighttime sequence. All images have resolution of 640 × 471.

4.2 Performance Evaluation Metrics

Towards this approach, E_G, *MI, EI, ISS*, and *HP* [34–39] metrics have been opted for the validation of the proposed technique.

Entropy (E_G): It is used to calculate the spatial information of an image. It tells about the richness of useful information. A higher entropy value signifies the excellent performance of the method. The mathematical representation is shown below.

$$E_G = \sum_{s=0}^{S-1} p_s \log p_s \tag{11}$$

where, 'S' represents the number of gray levels i.e., 256 and p_s contains the pixels with 's' gray values in the image.

MI: It tells about the transfer of the quantity of important useful information from the given input source images to the single fused image.

$$I = \sum_{f,a} p_{A'F'}(f,a) \log \frac{p_{A'F'}(a,f)}{p_{A'}(a)p_{F'}(f)} + \sum_{f,b} p_{B'F'}(f,b) \log \frac{p_{B'F'}(b,f)}{p_{B'}(b)p_{F'}(f)} \tag{12}$$

where, two source input images are described by A' and B', F' is the fused image. Joint histograms of the source input and fused output image are denoted by $p_{A'F'}$ and $p_{B'F'}$. Whereas, $p_{A'}$, $p_{B'}$, and $p_{F'}$ depicts the corresponding histograms of A', B', and F'.

Edge information: It calculates the transference of the visual as well as edge information from the two input source images to the fused image.

$$EI = \frac{\sum_{s=1}^{S} \sum_{t=1}^{T} (Q^{A'F'}(i,j)\, W^{A'}(i,j) + Q^{B'F'}(i,j)\, W^{B'}(i,j)}{\sum_{i=1}^{S} \sum_{j=1}^{T} (W^{A'}(i,j) + W^{B'}(i,j)} \tag{13}$$

where, A' and B' denotes the source input images, F' is a fused image. $W^{A'}(i,j)$ and $W^{B'}(i,j)$ are the weights of the pixels. $Q^{A'F'}(i,j)$ and $Q^{B'F'}(i,j)$ indicate the similarity between A' and B'. The location of an image is referred by (i, j).

Image structural similarity: It describes the amount of structural information preservation into the resultant single image. It tells about the similarity between the given input images with resultant single fused images.

$$SS = \sum c\left(w'\right)\left(\alpha\left(w'\right) Q_0\left(A', F'/w'\right) + \left(1 - \alpha\left(w'\right)\right) Q_0\left(B', F'/w'\right)\right) \tag{14}$$

where, A' and B' denotes input source images, F' is fused output images, $\alpha\left(w'\right)$ is a local weight which is $0 <= \alpha\left(w'\right) <= 1$. The value closer to 1 indicates more transfer of weights from an input image to the fused image, $c\left(w'\right)$ used for the computation of weights described and w' denotes the set of windows.

Human Perception: The IF method is dependent on the perception of humans which is calculated using human perception. Input as well as output images are filtered using a contrast sensitivity filter. Then, the calculation of the contrast preservation map is done. It is represented in Eq. (15).

$$HP = \alpha_{A'}(i,j)\, Q_{A'F'}(i,j) + \alpha_{B'}(i,j)\, Q_{B'F'}(i,j) \tag{15}$$

where $\alpha_{A'}(i,j)$ and $\alpha_{B'}(i,j)$ are a saliency map, $Q_{A'F'}(i,j)$ and $Q_{B'F'}(i,j)$ describe the similarity between input and resultant image.

All metrics values have ranges in the [0, 1] interval [29]. 0 indicates low-quality image whereas, 1 implies high-quality image.

4.3 Experimental Setup

In this study, Siamese CNN has been presented. It consisted of two branches having the same neural structure with the same weights for the extraction of the features of two different infrared/visible images. The network training has been done by using a framework of caffe [40]. Xavier algorithm is used for the weight's initialization of each convolutional layer.

Training has been done on 50,000 natural images derived from the ImageNet dataset [41]. Due to the lack of a labeled datasets, a Gaussian filter has been used to obtain the blurred version of the images. After that, for every blurred version of the image, 20 sets of 16 × 16 patch size are sampled from the input image. Thus, 1,000,000 sets of patches have been obtained. However, only about 10,000 images have been used. The softmax loss function has been used as an optimization objective whereas, minimization of loss function has been done using stochastic gradient descent. Weights have been updated by the rule given below [42].

$$\alpha_{i+1} = M\alpha_i - W * \gamma * w_i - \gamma * \frac{\partial L}{\partial W_i} \tag{16}$$

where, α is a momentum variable, i is an iteration index, M is a momentum set at 0.9, W is a weight decay that is set at −0.0005, L is the loss function and γ is a learning rate. Loss function derivative is denoted as $\frac{\partial L}{\partial W_i}$, w_i is a weight and Learning rate γ is chosen as 0.002. A higher learning rate has adverse effects on the calculated loss, whereas a smaller learning rate takes an increased epoch for system convergence. These values have been set after performing various experiments. Lastly, standardization has been obtained by balancing the dataset. Transformation and augmentations have been done on the infrared/visible images as shown below.

- Random flipping: both horizontal and vertical flipping is done.
- Rotation: both horizontal as well as vertical rotation of images is done by 90° and 180°.
- Gaussian filter: blurring of images are obtained for noise smoothening.

4.4 Subjective Visibility

The fusion result on the six different sets of infrared/visible images has been attained. Based on the fused images, it can be observed that infrared images have apparent objects and visible images have an obvious background. The techniques such as GF, DL, RW, PCA, and ADKL failed in retaining the objects presented in the images well.

From Fig. 3, it can be noted that in Figs. 3I–3III, the fused images produced by DL, PCA, and ADKL are low-intensity images, hence, not able to keep the intensities of the object information. They contained blurriness and artifacts in the images as shown by the area in red

boxes. Thus, unclear and poor quality of images have been obtained. The visual quality of images obtained from RW and GF methods are worst because there is information loss and the upper right corner of these images seemed to be darker than the original image with some distortion too. The image produced by DCT is better as compared to above-discussed techniques but is also incapable to extract all the information. Thus, the proposed technique overcame these problems very well as shown in Figs. 3I–3III by producing images of enhanced quality.

It is evident from the Figs. 3IV–3VI that the fused image produced by the proposed technique contained more detailed information of the target image depicting image characteristics as well. By contrast, DL, ADKL, and DCT generated a noisy, blurred image with a poor quality of fused image. The DCT technique produced some distortion giving the distorted image. From the fused images of GF, RW, and PCA it can be analyzed that all the contents from the source images are not transferred to the resultant output images. On comparative analysis of the proposed technique with the other techniques, it has been analyzed that other techniques exhibited the loss of contrast, brightness, edges as well as incapable of fusing many types of features among two different heterogeneous images. Thermal radiations of the infrared and target object of the visible image has not been retained by these techniques and most of the information got damaged. Hence, the proposed technique has outperformed all other techniques by producing the better fused image.

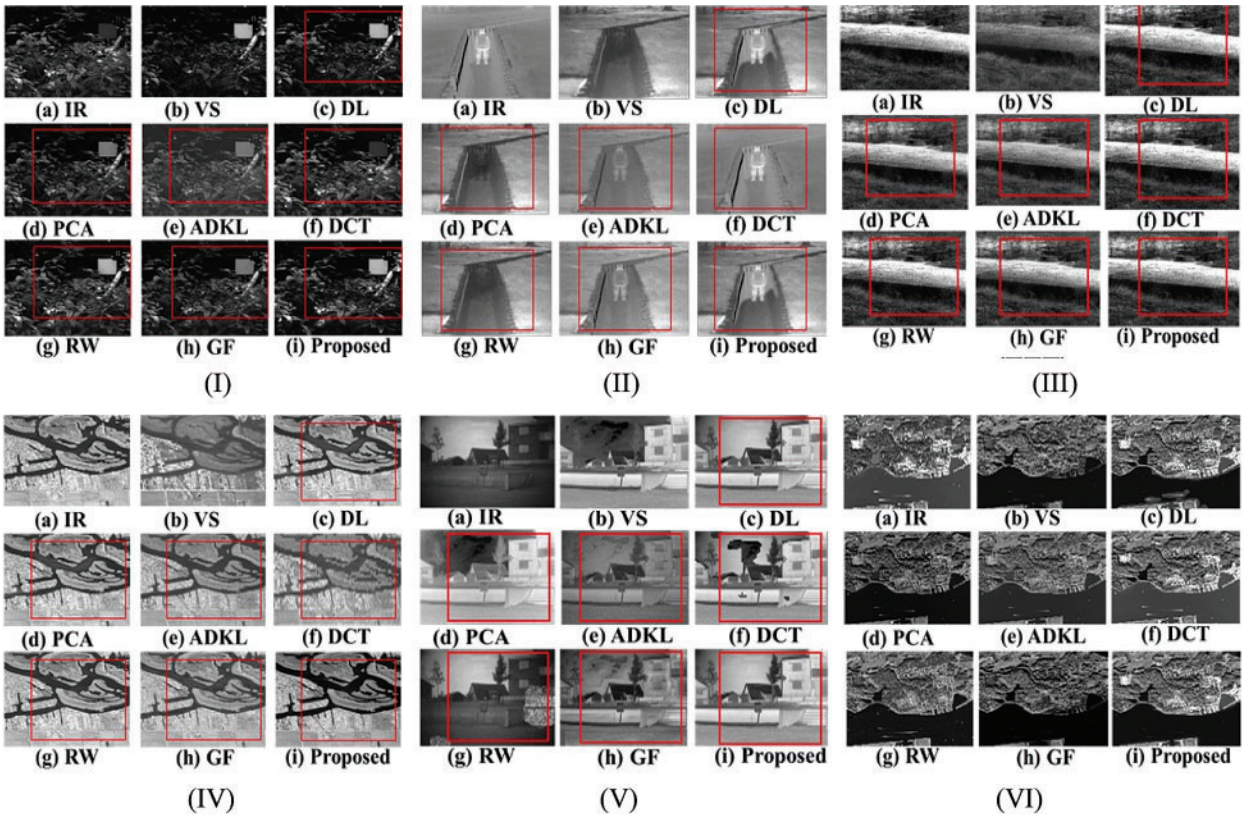

Figure 3: Qualitative fused images on 6 infrared/visible image pairs, (a)–(i) represents infrared image, visible image, fused output by DL, fused output by PCA, fused output by ADKL, fused output by DCT, fused output by RW, fused output by GF, and fused output by proposed technique in (I)–(VI)

4.5 Objective Visibility

For further illustrations of the fusion effects, five evaluation metrics such as MI, HP, ISS, E_G, and EI have been used. Higher values of metrics indicate the best quality of fused images. Its value ranges in the interval of 0 to 1 where 1 or more than 1 indicates the enhanced quality image [43–45]. Tab. 1, lists an average value of 78 sets of images. These values are compared on the basis of five evaluation metrics. It can be clearly observed that the proposed technique has fused the images with more MI, that is, more information transfer from the given input images to the resultant single fused image. It has also achieved the highest entropy which indicates that it has consisted of more spatial information of the given source images. On the contrary, ADKL has attained the lowest entropy which implies the ineffectiveness of this technique and less information has been transferred. Furthermore, the proposed technique also attained the highest values in terms of EI and HP that validate the fused image containing better visual edges and sharpening of the image. The proposed technique has also given the best ISS value near 1, which shows superiority in comparison to other existing techniques.

Therefore, from the above discussions, it can be concluded that the proposed technique attained the highest values in terms of every metric as shown in bold in Tab. 1. Hence, it has outperformed the other traditional infrared/visible image fusion techniques.

Table 1: Average comparison of metrics values for 78 sets of images

Fusion methods	MI	E_G	EI	ISS	HP
DCT	0.7318	0.7202	0.6099	0.5543	0.5472
PCA	0.6111	0.7040	0.6370	0.6728	0.5304
RW	0.7355	0.7994	0.5647	0.7421	0.5634
GF	0.3850	0.7689	0.6174	0.7510	0.5232
CNN	0.6490	0.7962	0.5015	0.8109	0.5115
ADKL	0.6709	0.6790	0.6180	0.6994	0.6978
Proposed	**0.9597**	**0.8429**	**0.6444**	**0.9709**	**0.6527**

5 Conclusion and Future Directions

This paper designed an infrared/visible image fusion technique based on the fuzzification and convolutional neural network. The main goal of this study is to solve the issue regarding the maintenance of thermal radiation features in the pre-existing IR/VS based methods. Therefore, benefits of two theories have been taken with the integration of FS and CNN to devise a new strong and adaptable technique into a single scheme. The proposed technique firstly retained the details of the thermal radiation of the infrared images, whereas simultaneously accumulated the visibility in the visible image. Therefore, correct target location can be observed which further helped in the processing and also vital for increasing precision and focused ability of the output image. This technique dealt with 78 sets of infrared/visible images. Furthermore, high quality and enhanced image has been produced even under bad illumination and varied expressions. The main goal behind this work is the designing of the advanced automatic technique to obtain the fused image containing contour, brightness, and texture information between IR/VS images to illustrate clear target features of the infrared image while distinctly visible background which will be further helpful in the military surveillance and object detection. The subjective, as well as

objective evaluation, indicated that the proposed technique has given a higher performance in comparison to the existing techniques in feature extraction and information gathering.

In the future, we intend on the optimization of the developed technique with the hybridization of the neuro fuzzy and CNN. Moreover, this technique can be more generalized for the fusion of more than two images at the same time by adapting the convolutional operations. Also, we intend to extend this research in other domains as well.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] J. Ma, P. Liang, W. Yu., C. Chen, X. Guo *et al.,* "Infrared and visible image fusion via detail preserving adversarial learning," *Information Fusion*, vol. 54, no. 1, pp. 85–98, 2020.

[2] G. Li, Y. Lin and X. Qu, "An infrared and visible image fusion method based on multi-scale transformation and norm optimization," *Information Fusion*, vol. 71, no. 1, pp. 109–129, 2021.

[3] J. Ma, Y. Ma and C. Li, "Infrared and visible image fusion methods and applications: A survey," *Information Fusion*, vol. 45, no. 1, pp. 153–178, 2019.

[4] J. Ma, Y. Ma and C. Li, "Infrared and visible image fusion methods and applications: A survey," *Information Fusion*, vol. 45, no. 1, pp. 153–178, 2019.

[5] X. Bai, "Morphological center operator based infrared and visible image fusion through correlation coefficient," *Infrared Physics & Technology*, vol. 76, no. 1, pp. 546–554, 2016.

[6] A. Toet, "Image fusion by a ratio of low-pass pyramid," *Pattern Recognition Letters*, vol. 9, no. 1, pp. 245–253, 1989.

[7] Y. A. Phamila and R. Amutha, "Discrete cosine transform based fusion of multi-focus images for visual sensor networks," *Signal Processing*, vol. 95, no. 1, pp. 61–70, 2014.

[8] J. A. Stark, "Adaptive image contrast enhancement using generalizations of histogram equalization," *IEEE Transactions on Image Processing*, vol. 9, no. 5, pp. 889–896, 2000.

[9] F. Chen, J. Zhang, J. Cai, T. Xu, G. Lu *et al.*, "Infrared image adaptive enhancement guided by energy of gradient transformation and multiscale image fusion," *Applied Sciences*, vol. 10, no. 18, pp. 6262–6292, 2021.

[10] D. P. Bavirisetti and R. Dhuli, "Fusion of infrared and visible sensor images based on anisotropic diffusion and krhunen-loeve transform," *IEEE Sensors Journal*, vol. 16, no. 1, pp. 203–209, 2015.

[11] Y. Liu, X. Chen, R. K. Ward and Z. J. Wang, "Image fusion with convolutional sparse representation," *IEEE Signal Processing Letters*, vol. 23, no. 12, pp. 1882–1886, 2016.

[12] C. Liu and W. Ding, "Variational model for infrared and visible light image fusion with saliency preservation," *Journal of Electronic Imaging*, vol. 28, no. 2, pp. 023023–123037, 2019.

[13] A. L. D. Cunha, J. Zhou and M. N. Do, "The nonsubsampled contourlet transform: Theory, design, and applications," *IEEE Transactions on Image Processing*, vol. 15, no. 10, pp. 3089–3101, 2006.

[14] H. Li, H. Qiu, Z. Yu and Y. Zhang, "Infrared and visible image fusion scheme based on NSCT and low-level visual features," *Infrared Physics & Technology*, vol. 76, no. 1, pp. 174–184, 2016.

[15] Y. Yang, Y. Que, S. Huang and P. Lin, "Multiple visual features measurement with gradient domain guided filtering for multisensor image fusion," *IEEE Transactions on Instrumentation and Measurement*, vol. 66, no. 4, pp. 691–703, 2017.

[16] J. Ma, Z. Zhou, B. Wang, L. Miao and H. Zong, "Multi-focus image fusion using boosted random walks-based algorithm with two-scale focus maps," *Neurocomputing*, vol. 335, no. 1, pp. 9–20, 2019.

[17] H. R. Shahdoosti and H. Ghassemian, "Combining the spectral PCA and spatial PCA fusion methods by an optimal filter," *Information Fusion*, vol. 27, no. 1, pp. 150–160, 2016.

[18] Y. Liu, X. Chen, H. Peng and Z. Wang, "Multi-focus image fusion with a deep convolutional neural network," *Information Fusion*, vol. 36, no. 1, pp. 191–207, 2017.
[19] H. Li and X. J. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2614–2623, 2018.
[20] H. Li, X. J. Wu and J. Kittler, "Infrared and visible image fusion using a deep learning framework," in *Proc. of 24th Int. Conf. on Pattern Recognition*, Beijing, China, pp. 2705–2710, 2018.
[21] J. Ma, W. Yu, P. Liang, C. Li and J. Jiang, "FusionGAN: A generative adversarial network for infrared and visible image fusion," *Information Fusion*, vol. 48, no. 1, pp. 11–26, 2019.
[22] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao *et al.,* "IFCNN: A general image fusion framework based on convolutional neural network," *Information Fusion*, vol. 54, no. 1, pp. 99–118, 2020.
[23] H. Xu, J. Ma, J. Jiang, X. Guo and H. Ling, "U2Fusion: A unified unsupervised image fusion network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, no. 1, pp. 1–17, 2020.
[24] G. Huang, Z. Liu, G. Pleiss, L. V. D. Maaten and K. Weinberger, "Convolutional networks with dense connectivity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, no. 1, pp. 1–12, 2019.
[25] F. Zhao, W. Zhao, L. Yao and Y. Liu, "Self-supervised feature adaption for infrared and visible image fusion," *Information Fusion*, vol. 1, no. 1, pp. 1–33, 2021.
[26] V. P. Ananthi and P. Balasubramaniam, "Image fusion using interval-valued intuitionistic fuzzy sets," *International Journal of Image and Data Fusion*, vol. 6, no. 3, pp. 249–269, 2015.
[27] H. Cai, L. Zhuo, X. Chen and W. Zhang, "Infrared and visible image fusion based on BEMSD and improved fuzzy set," *Infrared Physics & Technology*, vol. 98, no. 1, pp. 201–211, 2019.
[28] K. Zhang, Y. Huang, X. Yuan, H. Ma, C. Zhao *et al.,* "Infrared and visible image fusion based on intuitionistic fuzzy sets," *Infrared Physics & Technology*, vol. 105, no. 1, pp. 103124–103143, 2020.
[29] L. A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, no. 3, pp. 338–353, 1965.
[30] T. Yang and L. B. Yang, "Fuzzy cellular neural network: A new paradigm for image processing," *International Journal of Circuit Theory and Applications*, vol. 25, no. 6, pp. 469–481, 1997.
[31] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
[32] S. Lin, Z. Han, D. Li, J. J.Zeng, X. Yang *et al.,* "Integrating model-and data-driven methods for synchronous adaptive multi-band image fusion," *Information Fusion*, vol. 54, no. 1, pp. 145–205, 2020.
[33] A. González, Z. Fang, Y. Socarras, J. Serrat, D. Vázquez *et al.,* "Pedestrian detection at day/night time with visible and FIR cameras: A comparison," *Sensors*, vol. 16, no. 6, pp. 820–837, 2016.
[34] S. Bhat and D. Koundal, "Multi-focus image fusion using neutrosophic based wavelet transform," *Applied Soft Computing*, vol. 106, no. 1, pp. 107307–107339, 2021.
[35] S. Bhat and D. Koundal, "Multi-focus image fusion techniques: A survey," *Artificial Intelligence Review*, vol. 1, no. 1, pp. 1–53, 2021.
[36] H. Kaur, D. Koundal and V. Kadyan, "Image fusion techniques: A survey," *Archives of Computational Methods in Engineering*, vol. 1, no. 1, pp. 1–23, 2021.
[37] S. Bhat and D. Koundal, "Multi-focus image fusion: Quantitative and qualitative comparative analysis," in *Proc. of ICRIC*, pp. 533–542, 2019. https://doi.org/10.1007/978-3-030-29407-6.
[38] H. Kaur, D. Koundal and V. Kadyan, "Multi modal image fusion: Comparative analysis," in *Proc. of Int. Conf. on Communication and Signal Processing*, Chennai, India, pp. 0758–0761, 2019.
[39] K. B. Chitkara, B. Sharma and D. Koundal, "Comparative analysis of image fusion methods," in *Proc. of 6th Int. Conf. on Computing for Sustainable Global Development (INDIACom)*, Delhi, India, pp. 535–541, 2019.
[40] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long *et al.,* "Caffe: Convolutional architecture for fast feature embedding," in *Proc. of the 22nd ACM Int. Conf. on Multimedia*, Orlando, Florida, USA, pp. 675–678, 2014.
[41] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, no. 1, pp. 1097–1105, 2012.

[42] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. of the Thirteenth Int. Conf. on Artificial Intelligence and Statistics*, Sardinia, Italy, pp. 249–256, 2010.

[43] H. Kaur, D. Koundal, V. Kadyan, N. Kaur, K. Polat *et al.,* "Automated multimodal image fusion for brain tumor detection," *Journal of Artificial Intelligence and Systems*, vol. 3, no. 1, pp. 68–82, 2021.

[44] S. Basheer, S. Bhatia and S. B. Sakri, "Computational modeling of dementia prediction using deep neural network: Analysis on OASIS dataset," *IEEE Access*, vol. 9, no. 1, pp. 42449–42462, 2021.

[45] K. Dev, S. A. Khowaja, A. S. Bist, V. Saini and S. Bhatia, "Triage of potential COVID-19 patients from chest X-ray images using hierarchical convolutional networks," *Neural Computing and Applications*, vol. 1, no. 1, pp. 1–6, 2021.