On Design of Image Fusion and Object Detection Models for Thermal Video Surveillance

Thesis submitted for the award of the Degree
of

Doctor of Philosophy

in the Department of Electrical Engineering

by

Manoj Kumar Panda
(2018REE0050)

Under the supervision of

Dr. Badri Narayan Subudhi

भारतीय प्रौद्योगिकी
संस्थान जम्मू
INDIAN INSTITUTE OF
TECHNOLOGY JAMMU

Indian Institute of Technology Jammu
Jammu 181221

April 2022

Declaration

I hereby declare that the matter embodied in this thesis entitled **"On Design of Image Fusion and Object Detection Models for Thermal Video Surveillance"** is the result of investigations carried out by me in the Department of Electrical Engineering, Indian Institute of Technology Jammu, India, under the supervision of **Dr. Badri Narayan Subudhi** (IIT Jammu) and it has not been submitted elsewhere for the award of any degree or diploma, membership etc. In keeping with the general practice in reporting scientific observations, due acknowledgements have been made whenever the work described is based on the findings of other investigators. Any omission that might have occurred due to oversight or error in judgment is regretted. A complete bibliography of the books and journals referred in this thesis is given at the end of the thesis.

April 2022
Indian Institute of Technology Jammu

Manoj Kumar Panda
(2018REE0050)

Dedicated to my Parents,

My beloved Wife and Son

Abstract

Recently, thermal surveillance has gotten much attention due to its possible applications in military and naval technologies. However, due to the developed technology and cheaper price of thermal sensors, they are put forward for various applications: agriculture, food industry, night surveillance, building inspection, gas detection, industrial safety, etc. The thermal sensor captures heat radiated by the object and is independent of the natural source of energy. Therefore, automatic surveillance can be performed using the thermal camera on a 24-hour basis and at any environmental condition. However, the performance of the thermal surveillance is challenging due to the distinct characteristics of the thermal image: low resolution (or) missing information, low signal-to-noise ratio, lack of structure such as shape and textural information, lack of color information, and low contrast, etc. Thus, the visual contents of the thermal image are poorer and make it difficult to detect the moving object present in the thermal scene. Hence, there is a need for designing of technologies to enhance the perceivable information and use the same to detect moving objects.

The objective of the thesis is to investigate and analyze, both theoretically and empirically, by developing some new algorithms to improve the visual contents and detect the moving objects in the thermal video scene for surveillance systems.

To enhance the visual contents in the thermal image, we have proposed two-pixel level (visual and thermal) image fusion techniques: fuzzy edge preserving intensity variation approach and a weighted combination of maximum and minimum value selection strategy. In the proposed fuzzy edge preserving intensity variation approach, we have explored the maximum selection strategy and fuzzy edge preservation mechanism to generate the high contrast fused image with significant edge details. Further, in the proposed weighted combination of maximum and minimum value selection strategy, we have investigated the maximum selection strategy, the minimum selection strategy, and the weighted combination mechanism among the source images to generate the fused image with essential details and reduced artifacts. However, it is observed that the proposed pixel-level image fusion schemes have introduced many isolated points in the fused image, which degraded the quality of the fused images.

Further, we have proposed two feature level (visual and thermal) image fusion schemes: integration of bi-dimensional empirical mode decomposition with two streams VGG-16

and non-subsampled contourlet transform induced two streams ResNet-50 network. In the proposed integration of bi-dimensional empirical mode decomposition with two streams VGG-16 technique, the bi-dimensional empirical mode decomposition (BEMD) mechanism integrated with VGG-16 architecture is proposed to preserve the deep features at various levels. The proposed deep multi-level fusion strategy produces the fused image with complementary details. Again, in the non-subsampled contourlet transform induced two streams ResNet-50 network technique, the non-subsampled contourlet transform (NSCT) mechanism induced with the ResNet-50 network is exploited the deep directional features at low-frequency and high-frequency bands. The proposed fusion strategy generates the fused image with reduced artifacts. The fused image sequences are then used for object detection.

In the next stage of the thesis, We have proposed a kernel induced possibilistic fuzzy associate background subtraction technique for moving object detection. In the proposed kernel induced possibilistic fuzzy associate background subtraction technique, the use of the induced kernel function projects the low dimensional data into a higher dimensional space and the use of the possibilistic function constructs a robust background model based on the density of the data in the temporal domain avoiding the noisy and outlier points. Further, it is observed that detecting the accurate shape of the moving object is quite difficult and again a thermal video contains high uncertainty due to low resolution and high noise.

In this context, we have proposed two multi-scale deep learning architectures for moving object detection: modified ResNet-152 network with hybrid pyramidal pooling and multi-scale contrast preserving deep learning for moving objects detection. In the proposed modified ResNet-152 network with a hybrid pyramidal pooling technique, a modified ResNet-152 network is induced on the multi-scale features extraction (MFE) block to enhance the feature learning capabilities that preserve sparse and dense deep features. The proposed decoder network precisely projects the feature-level into pixel-level. Again, in the proposed multi-scale contrast preserving deep learning architecture, the proposed encoder network with multi-scale contrast preservation (MSCP) block can retain the contrast details of the in-depth features. The proposed decoder network accurately projects the extracted features at different layers into pixel-level.

The efficiency of the proposed techniques is corroborated by testing them on differ-

ent benchmark databases. The performances of the proposed methods are evaluated by considering different competitive state-of-the-art techniques using relevant quantitative measures.

Acknowledgements

The journey of the pursuit of my doctorate degree has been nothing short of enthralling in every aspect. I now understand what it truly means when people say that a PhD changes a person in ways it is hard to imagine. This journey has meant so much more to me than merely advancing my academic ventures. It has changed my outlook on life as a whole. I find myself more adept at tackling the challenges coming my way. I find much more tools at my disposal than I ever thought I had possess. My mind feels more resilient to outcomes that may not always be in my favour. In short, I am a better version of what I used to be four years ago.

The credits for a transformation of this magnitude can hardly be given to a sole individual. My case is no different. It seems impossible to recount every person who contributed to the culmination of my degree, but I shall try my best to give credit where it is due. I am wholeheartedly indebted to my supervisor, Dr. Badri Narayan Subudhi, who graciously accepted me under his guidance and has been a beacon of light throughout this journey. None of my work would have been possible if not for his decades of expertise in image and video processing, and Machine Learning. Academic mentoring aside, he has been an unparalleled pillar of emotional support and motivation through the ups and downs over these years.

I am beholden to the members of my SRC committee, Dr. Karan Nathwani, and Dr. Ashok Bera, for painstakingly monitoring the intricacies of developments in my research and ironing out the wrinkles in my thesis. This work would not have been possible without their constant and insightful guidance.

I would like to acknowledge the efforts of my collaborators, Dr. T. Veerakumar, Dr. Vinit Jakhetiya, and Prof. Manoj Singh Gaur, who have been extremely accommodating and patient with me throughout our research ventures together. I owe it to them to have had the honour of being recognized in several world-acclaimed journals and conferences.

I am blessed to have received the help of my seniors, especially Dr. Deepak Kumar Rout during this journey. I would also like to thank Dr. Priyadarshi Kanungo, C. V. Raman Global University for the mental support and encouragement during the research work.

A very special note of thanks goes to my colleagues with whom I have had the pleasure of sharing this journey very intimately. I would like to thank Mr. Himanshu Singh, Ms.

Meghna and Mr. Tirupati Pallewed for being the invaluable help whenever I needed it. I would also like to thank Mr. Pawan Kumar, Mr. Ritujoy Biswas, Mr. Murtiza Ali, Mr. Rantu Buragohain, and Mr. Pradosh Kumar Hota for supporting and encouraging me whenever I hit a roadblock or when times were not in my favour. I owe it to them to have kept a fun, light-hearted and competitive environment at our workplace.

I am forever indebted to my wife, Snigdha, and my son, Sashanka, for giving their unflagging love and care during the journey of my doctoral degree. None of it would have been possible if not for their constant love, support and patience. I would like to thank my parents, and my brothers, for always having faith in me and encouraging me throughout this journey. It would not be wrong to say that my family has made me able to complete my thesis. This is why; I would like to dedicate this work to them.

I would also like to thank our Head of the Department, Dr. Ankit Dubey, for always being a helpful and supportive mentor, and always helping me in administrative affairs. I would like to thank our Director, Prof. Manoj Singh Gaur and the entire IIT Jammu family at large for providing me with the opportunity to work, learn and grow in this prestigious institution.

Manoj Kumar Panda

Contents

List of Figures

List of Tables

List of Symbols

Symbol	Description
A_s^i	Activity level map
$\tilde{A}_s^i$	Action level map
B	Fuzzy set
β	Weight values for the coefficients of salient features
c_l	Velocity of light
EPI_a	Average edge preservation index
e^{I_1}	Edge strength of the visible image
e^{I_2}	Edge strength of the IR image
e^F	Edge strength of the fused image
F	Fused image
FMI_{dct}	Mutual information for the discrete cosine features
FMI_w	Mutual information for the wavelet features
F_e	Exponential fuzzifiers
F_d	Denominational fuzzifiers
$\mathcal{F}$	Deep feature maps from the encoder network
γ_j	Spread of the j^{th} background type
h	Planck's constant
I_1	Visible image
I_2	IR image
IMF_{sn}	n numbers of intrinsic mode function extracted from s^{th} source images
k_B	Boltzmann's constant
$K(.)$	Gaussian kernel function
λ	Spread function
λ_r	Wavelength of the radiation
$MI(I_s; F)$	Mutual information among the source and fused images
$\bar{\mu}_i$	Mean of the source image patches
$\bar{\mu}_f$	Mean of the fused image patches
μ	Membership function
N_{abf}	Amount of artifacts added during the fusion process
P_l	Normalized histogram
$P_I(i)$	Marginal probability of source image
$P_F(f)$	Marginal probability of fused image
$P_{I,F}(i, f)$	Joint probability of source and fused image

Symbol	Description
p_{mn}	Property plane
P_t	Input of the t^{th} residual block
Φ	Kernel function
ψ_i	Local change detection output image
$\bar{Q}^{I_1F}$	Quantity of data that propagates from the visible to fused image
$\bar{Q}^{I_2F}$	Quantity of data that propagates from the IR to fused image
Q_t	Possibilistic fuzzy associated cost function
$\mathcal{Q}_t$	Output of the t^{th} residual block
$\Re(\cdot)$	ReLU function
r	Fuzzification parameter
R_d	Dilation rate
$SSIM_a$	Average structural similarity index
σ_i	Standard deviation of the source image
σ_f	Standard deviation of the fused image
σ_{if}	Covariance
sd	Stride
Θ_s^i	Deep features extracted by the deep neural network network
T	Temperature
v_j	Mode corresponding to each background type
$w \times w$	Size of center sliding window
$\mathcal{W}_t$	Set of weights and biases
W_s^i	Action weight map
$\tilde{W}_s^i$	Intermediate weight map
(x, y)	Pixel location
$\mathcal{X}$	Output of the pyramidal pooling architecture
$\bar{\mathcal{X}}_u$	Contrast details
$\mathcal{Y}$	Output of the multi-scale features extraction block
$\bar{\mathcal{Y}}$	Output of the multi-scale contrast preservation block

List of Abbreviations

Abbreviation	Description
AIT	Airplane in Trees
ACC	Accuracy
BGS	Background Subtraction
BEMD	Bi-dimensional Empirical Mode Decomposition
BBAO	Block-based Average Operator
BBA	Block-based Average
BR	Bunker
BCEL	Binary Cross-entropy Loss
CSR	Convolutional Sparse Representation
CS-MCA	Convolutional Sparsity Based Morphological Component Analysis
CNN	Convolutional Neural Network
CP	Contrast Pyramid
CM	Camp
CN	Contrast Normalization
Cascade CNN	Cascade Convolutional Neural Network
DFM	Detail Feature Map
DL	Deep Learning
DFBs	Directional Filter Banks
DNN	Deep Neural Network
EN	Entropy
EPI	Edge Preservation Index
EMD	Empirical Mode Decomposition
EB	Encoder Block
FIR	Far-infrared
FSD	Filter Subtract Decimate Pyramid
FusionGAN	Fusion Based on Generative Adversarial Network
FP	False Positive
FN	False Negative
FPR	False-positive Rate
FNR	False-negative Rate
FLIR	Forward Looking Infrared
GP	Gradient Pyramid
GAP	Global Average Pooling
GPU	Graphics Processing Unit
IR	Infrared

Abbreviation	Description
IFM	Intermediate Feature Map
IMF	Intrinsic Mode Function
IFCNN	Image Fusion Based on CNN
KP1123	Kaptein 1123
KP1654	Kaptein 1654
LWIR	Long-wavelength Infrared
LP	Laplacian Pyramid
LE	Lake
MST	Multi-scale Transform
MP	Morphological Pyramid
MBBA	Moving Block Based Average
MFE	Multi-scale Features Extraction
MSCP	Multi-scale Contrast Preservation
MI	Mutual Information
MFT	Man in Front of House
MID	Men in Doorway
ME	Marne
MV1	Movie 1
MV18	Movie 18
MCC	Matthews Correlation Co-efficient
NIR	Near-infrared
NSCT	Non-subsampled Contourlet Transform
NSPFBs	Non-subsampled Pyramid Filter Banks
NSDFBs	Non-subsampled Directional Filter Banks
pdf	Probability Density Function
PPA	Pyramidal Pooling Architecture
PWC	Percentage of Wrong Classification
RP	Ratio of Low-pass Pyramid
RP-SR	Ratio of Low-pass Pyramid Sparse Representation
RGB	Red, Green, Blue
ROI	Region of Interest
ReLU	Rectified Linear Unit
RBO	Residual Block Operation
RFN	Residual Fusion Network
RMSProp	Root Mean Squared Propagation
SWIR	Short-wavelength Infrared
SOTA	State-of-the-Art
SFM	Salient Feature Map

Abbreviation	Description
SAD	Sum of the Absolute Difference
SSIM	Structural Similarity Index
SLD	Saliency Detection
ST	Street
SP	Sandpath
SBS	Soldier Behind Smoke
SIT	Soldier in Trench
SNR	Signal to Noise Ratio
SD	Spatial Dropout
Tconvs	Transposed Convolution Layers
TU-VDN	Tripura University Video Dataset at Night Time
TC	Transposed Convolution
TP	True Positive
TN	True Negative

Chapter 1

Introduction and Scope of the Thesis

1.1 Introduction

In the past few decades, research and development in automatic vision-based surveillance systems have been getting its popularity because of their massive real life applications. It is a process of detecting moving objects followed by tracking of the same from a video scene. For the surveillance system, the visual sensor is an essential component. The visual sensor can capture the image in grey scale or RGB plane with detailed textural information and better spatial resolution. However, these kinds of images are immensely afflicted by variations in illumination. Furthermore, the visual sensor is unable to provide a better accuracy during night time due to low visibility. Therefore, to surpass the said limitations thermal sensors are widely used for night time surveillance. Thermal sensors capture the heat emitted by the object and are hence independent of the external sources. Based on their capability to sense the radiated heat, thermal sensors are found to be used as two varieties: active and passive thermal sensors. An active sensor emits infrared radiation and captures the radiation reflected by an object. Generally, the active sensors capture the radiation of both the near-infrared electromagnetic and the visible spectrum and less reliant on illumination. However, passive thermal sensors are preferred in many applications as they captures heat radiated by the object, breaking the dependency on external sources of energy. Thermal radiation is not affected by adverse environmental conditions but generally has ambiguous textural information and poor resolution. Therefore, most of the time it is preferred to use both visible and thermal sensors simultaneously for the

surveillance system instead of using either visible sensor or thermal sensor individually.

1.2 Visible Imaging

An image represents the real world scene in a 2D plane. A camera captures the light energy reflected from objects on the camera's view. The world we reside in is 3D, but the camera focuses the rays falling on the screen to a 2D image plane. Mathematically, an image can be represented as a two-dimensional function $I(x, y)$, where (x, y) denotes the spatial coordinates, and I indicates the intensity value at that point (x, y), generally termed as pixel. A graphical illustration of the conventional image formation model is shown in Figure 1.1. The images formation model has three main components: light source, camera, and target object. The source generates a ray of photons if it is in the visible spectrum. Sun is the natural source of light, and it acts as a source of illumination for the natural photography. The light from the source incident on the target and reflected light is allowed to fall on the lens of a camera based on the reflectance coefficients of the object. The light rays falls on the lens and projected to the imaging sensor generates an image on the imaging plane. Sampling and quantization strategies are applied to represent the image in the digital domain. In the digital image, the intensity value of each pixel is represented by b bits binary digits. Generally, for a 8 bits system, the pixel intensity values range from 0 to 255. Further, several image processing task can be made to extract meaningful information from an image: image enhancement, image restoration, image compression, image segmentation, and object recognition.

1.3 Thermal Imaging

Thermal imaging is an approach of converting thermal radiation information into visual information representing the spatial distribution of temperature variation in a scene captured by a thermal camera. The infrared (IR) sensor captures the heat emitted by the objects, and the infrared lens focuses the radiation information on the focal plane array. The IR-sensitive detector coincides with the focal plane array using the photoelectric effects to generate the electrical signal and is processed by the image processor to generate the image. Here, the intensity level of the image is proportional to the object's temperature. The IR detector may be cooled, or uncooled based on the materials used

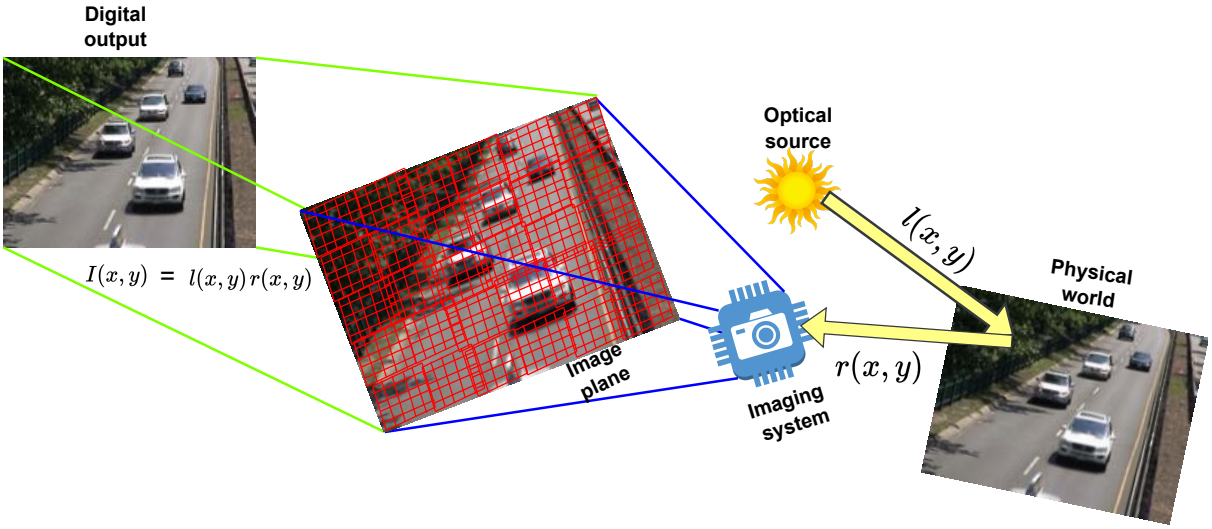


Figure 1.1: Visible image formation model.

in the detector and the sensors use. The image formation model by the thermal sensor is presented in Figure 1.2. The wavelength spectrum of infrared radiation lies between

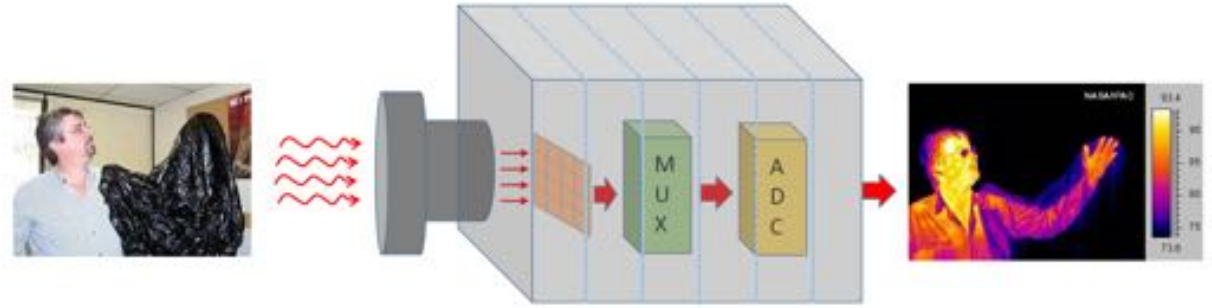


Figure 1.2: Thermal image formation model.

the visible light range and the microwave range of the electromagnetic spectrum. The infrared spectrum is divided into five categories: near-infrared (NIR), short-wavelength infrared (SWIR), mid-wavelength infrared (MWIR), long-wavelength infrared (LWIR), and far-infrared (FIR) as represented in Figure 1.3. Generally, the MWIR and LWIR infrared spectrum are known as thermal infrared as the objects emits the thermal radiation in this spectral range. Due to presence of various particles and gasses in the atmosphere it can allow the certain range of thermal radiation spectrum and attenuate the others. Figure 1.4, demonstrate the atmospheric windows for thermal radiation transmission and attenuate the thermal radiation because of the absorbing molecule in the atmosphere. Radiation from an object at the temperature T can be determined by Planck's radiation

law and can be given as,

$$I(\lambda_r, T) = \frac{2\pi h c_l^2}{\lambda_r^5 (e^{hc_l/\lambda_r k_B T} - 1)}, \tag{1.1}$$

where λ_r denotes the wavelength of the radiation, h is the Planck's constant, c_l represents the velocity of light, and k_B is the the Boltzmann's constant. Figure 1.5 depicts few real life infrared images at various conditions.

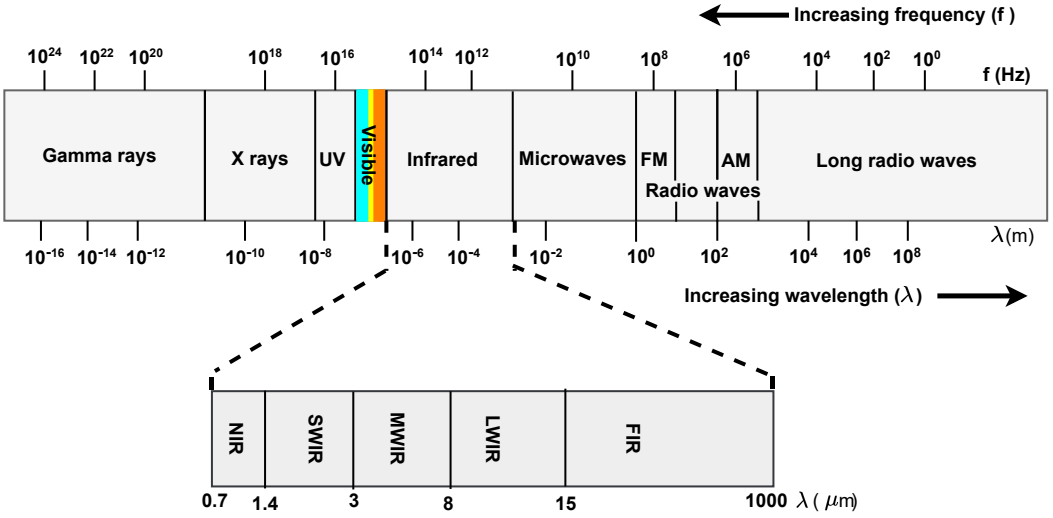


Figure 1.3: The electromagnetic spectrum with various division of infrared spectrum.

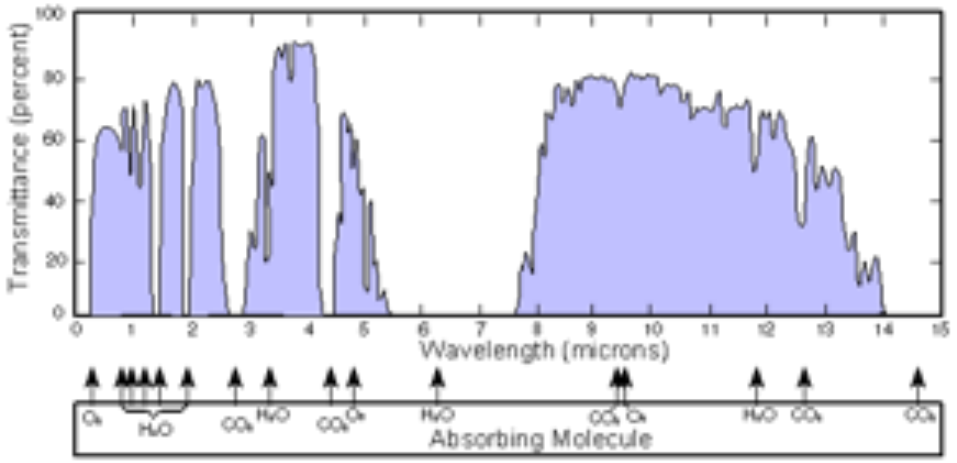


Figure 1.4: Transmitting and attenuated region of infrared radiation through atmosphere [1].

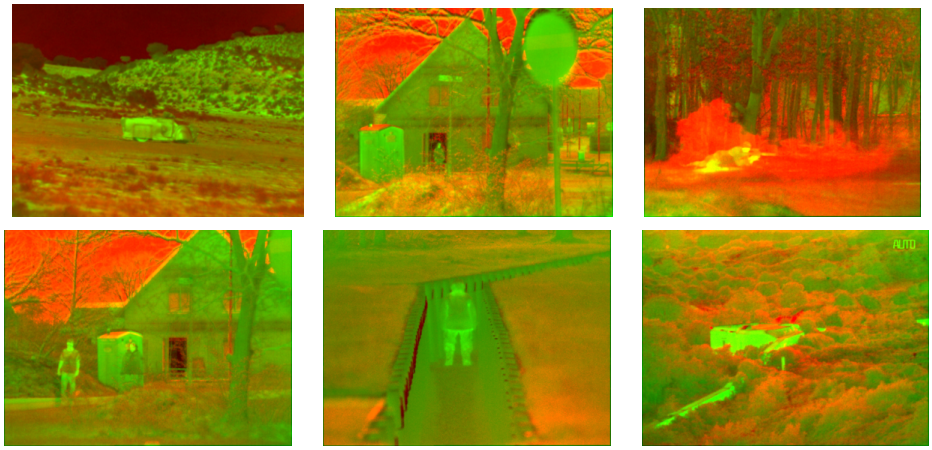


Figure 1.5: Examples of real life infrared images [2].

1.3.1 Thermal Sensors

Generally, two types of detectors are used in the thermal sensing: photon detector or thermal detector [3]. The photon detector transforms the infrared radiation into electric energy by varying the free charge concentration in a semiconductor. This type of detector generally works in the mid-wavelength infrared band and is highly sensitive to variation in the scene temperature. The photon detector-based infrared sensor generates a higher frame rate video. However, while capturing the image, the said sensor needs to be cooled to reduce the sensor noise. This cooling mechanism can be attained using liquid nitrogen or a cryocooler. Therefore, this system is not cost-effective.

A thermal detector transforms infrared radiation into thermal energy, which causes a rise in temperature at the detector. Then potential is developed across the sensor corresponding to the temperature. The thermal detector is an uncooled device that uses ferroelectric detectors or microbolometers. The ferroelectric detectors utilize the barium strontium titanate (BST) material, where minor variations in the scene temperature cause more apparent changes in electric polarization in the material. However, the images from the BST detector has low resolution and produces a Halo effect around the edges in the image. Therefore, the thermal detector based on microbolometers is popular, utilizing amorphous silicon and vanadium oxide materials. The thermal radiation of an object changes the property (electrical resistance) of the material; thus, potential get developed and further processed to generate an image. The thermal sensor based on microbolometers generates an image with better spatial resolution and reduced thermal noise.

Generally, the thermal images are represented as greyscale images with a bit depth of

8 or 16 bits. For better visibility of the thermal image, pseudo color is assigned during the image representation. The spatial resolution for the standard thermal sensor typically varies from 160×120 pixels to $1,280 \times 1,024$ pixels.

1.4 Automatic Surveillance System

In automatic surveillance, the system monitors people through cameras placed around different geographical locations. The far most objective of any surveillance system is to extract meaningful information from the video data captured from the surveillance cameras by detecting, and tracking the moving objects, and analyzing their behaviors. Surveillance may have different applications including: crime prevention, homeland security, traffic behavior analysis, risk prediction, indoor monitoring of elder persons and children, etc. Visual surveillance can be accomplished by using either a visible sensor or a thermal sensor. In the past few decades, visible sensors play an import role for surveillance system. However, such type of system fails in adversarial environmental condition, unable to capture the image in darkness or low light condition like night time. Hence, to make the surveillance system automated and facilitates continuously monitoring the objects on a 24-hour basis, a thermal sensor is preferred these days. The graphical illustration of the thermal surveillance system is shown in Figure 1.6. Thermal camera for surveillance is one of the prime areas of focus for several military and naval applications. But, due to the advanced technology and cheap availability of thermal sensors, they are put forward for many other applications: night surveillance [4], agriculture and food industry [5], building inspection [6], gas detection [7], industrial safety [8], etc.

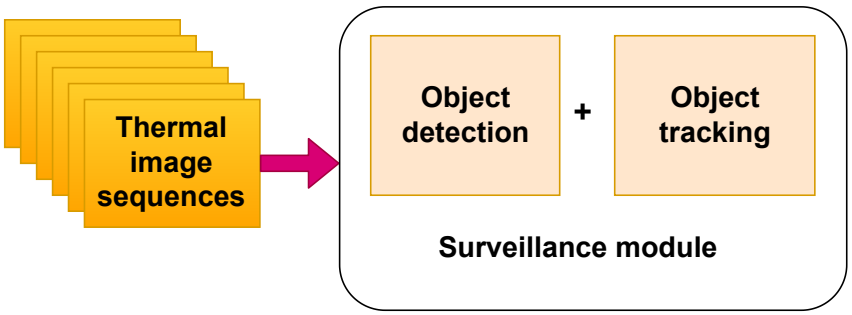
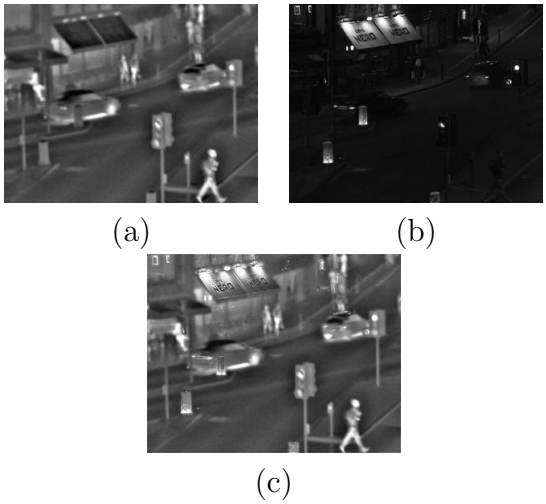


Figure 1.6: Block diagram of the thermal surveillance system.

1.4.1 Needs of Visual Enhancement by Image Fusion

The thermal sensor has captured the infrared radiation information emitted by the objects and is represented as bright pixels for hot objects [9]. Generally, the IR images provide information based on thermal radiation from the objects, usually characterized by pixel values where the targets are recognized but have insufficient background information. Also, the IR image has ambiguous textural information and poor resolution. Therefore, it is required to increase the perceivable information in the IR image, known as the visual enhancement. However, this will be more challenging as thermal images are low resolution and have poor textural information. Hence, most of SOTA techniques have explored the fusion of visible and thermal images for visual enhancement. Figure 1.7 (a) and (b) show an IR image and a visible image pair. Figure 1.7 (c) is the fused image obtained by using the image fusion technique that preserves complementary contents from the source images. From Figure 1.7 (c), it may be found that the fused image contains more information and is visually appealing. Therefore, in thermal video surveillance, visual enhancement is an important step before object detection as well as tracking of the same. Hence, the modified graphical illustration of the automatic video surveillance system is presented in Figure 1.8.



(a) (b)

(c)

Figure 1.7: An example of image fusion (a) infrared image, (b) visible image, and (c) fused image .

1.4.2 State-of-the-Art Techniques for Image Fusion

Image fusion is a technique of combining information from different images of the same scene into a single image [10]. The fused image may contain more important information

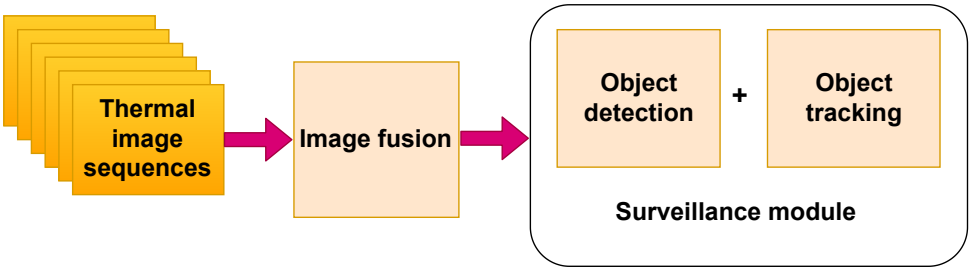


Figure 1.8: Block diagram of the modified thermal surveillance system.

that can help ensue processing or help in decision making. The important requirements for image fusion are extracting salient features from the different source images and using them in combination to produce the fused one with reduced artifacts. Many techniques have been proposed in the last few decades to extract the features from the sources and construct the fused image [11]. These fusion approaches are successfully used in different applications: military application, video surveillance, medical domains, etc, and found to be achieving a better performance.

For the vision based system, we can use either a individual imaging type (visible or IR) or both. The advantages of visible images usually have good spatial resolution and detailed textural information; thus they are applicable for human visual intuition. However, these kind of source images can be simply influenced by darkness and luminosity, disturbed weather conditions such as rain, fog, smoke, snow, etc. In-order to overcome these problems and to appropriately carry out object detection, several research articles suggest that the use of IR image for fusion with the visible one. Actually, an IR sensor captures the temperature emitted by target object. So it doesn't get effected by sudden change in environmental conditions [12] and provides suitable information of the objects at night time or at disturbed weather conditions. This typical capability of IR sensor facilitates continuously monitoring objects on a 24-hour basis and can detect hot objects. But the IR based system unable to handle information in a hot day as it provides a lot of hot areas which emits heat including the objects. Also thermal radiation of objects typically have poor texture and low resolution. Therefore, it is of great advantage to fuse the IR and visible images for vision system instead of using single type of image (visible or IR) for the surveillance system. By virtue, it is assumed that different source images are registered pixel by pixel [13]. IR and visible image fusion has widely been used in the domains of object tracking [14], object detection [15], object recognition [16], surveillance

[17], remote sensing [18], military [19], etc.

Depending on the applications and information representation, the SOTA of image fusion are categorized of three types: pixel level, feature level, and decision level [20].

1.4.2.1 Pixel-level Image Fusion Techniques

It is the process of directly combining the original information of the source images to generate the fused image, which is more informative than the source images [21]. During past decades, many techniques have been presented to report the fusion at pixel level. The simplest approach to the fusion of images is pixel-by-pixel averaging of the images from different sensors [22]. However, the main drawback of this technique is, the sources have similar effects in the fused image without considering their information contents. To overcome this problem, the frequently used techniques are multi-scale transform (MST). The MST depended techniques encompass the subsequent steps [20]. Initially, the images obtained by the various sensors are decomposed into several layers with different salient features. Then, by utilizing appropriate fusion rules different layers are fused. At the end, by utilizing the relating inverse MST on the fused layers, the final fused images are reconstructed. The most frequently used multi-scale transform-based techniques for fusion of images include: Laplacian pyramid (LP) [23], ratio of low-pass pyramid (RP) [24], contrast pyramid (CP) [25], filter-subtract-decimate pyramid (FSD) [26], gradient pyramid (GP) [26], and morphological pyramid (MP) [26], discrete wavelet transform (DWT) [27], shift-invariant discrete wavelet transform (SI-DWT) [27] curvelet transform (CVT) [28], and nonsubsampled contourlet transform (NSCT) [29]. However, the multi-scale transform-based techniques are trying to retain similar important information from both sources, which is not appropriate for IR and visible image fusion.

Li *et al.* [30] put forward a guided filter-based image fusion technique, where an average filter is used for two-scale decomposition. Then, a weighted average is used for the fusion of the base and the detailed parts. However, such methods are not able to retain the edges while smoothing images according to their scales, which provides a greater advantage to the fusion scheme. Bavirisetti and Dhuli [31] proposed an IR and visible image fusion technique based on a two-scale decomposition and saliency detection, where the mean and the median filters are used to obtain the base and the detailed components. Next, weight maps are obtained from the visual saliency. By combining the base, the detailed, and

the weight map components, the fused image is constructed. However, such an approach gives Halo effects around the edges in the fused images.

1.4.2.2 Feature-level Image Fusion Techniques

It is the process of incorporating the feature sets extracted from the source image to produce the fused image, which may contain more essential information than the source images [32]. During past decades, many techniques have been presented to report the fusion at the feature level. Zong *et al.* [33] proposed SR technique-based fusion method where the classification of image segments and learning numerous sub-dictionaries has been done using Histogram of Oriented Gradient (HOG) features. The resultant image is obtained using the l_1-norm and the maximum selection strategy. Being highly sensitive to mis-registration and having a limited capability to preserve the details are the major setbacks of these SR-based fusion techniques. In [34], the authors proposed a convolutional sparse representation (CSR) for fusing images, where the CSR-based method obtains deep features that are used for the fusion. However, this technique provides fused images containing ringing artifacts around the salient features. Liu *et al.* [35] proposed an image fusion approach of morphological component analysis based on convolutional sparsity (CS-MCA). By integrating morphological component analysis (MCA) and convolutional sparse representation (CSR) into a unified optimization network, this method achieves multi-component and global SRs of the source images. Here, the CSRs and texture components are generated by the CS-MCA model [35] using pre-learned dictionaries. The fused image is obtained by using the CSRs and the texture components of the source images. However, this fusion scheme is suited to the fusion of multi-focus images.

For the last few years, deep learning has been a lucrative tool for the extraction of deep features from source images, which are also used for the fusion of images at feature level. An IR and visible fusion of images with a convolutional neural network (CNN) is proposed in [36]. In this technique, the Laplacian pyramid decomposition is used for the source images. Again, a convolutional network is applied to obtain the weight maps from the source images and the Gaussian pyramid decomposition for the weight maps. Finally, inverse Laplacian transform and coefficient fusion are used to construct the fused image. However, the said technique cannot take full advantage of extracted features. A CNN-based image fusion scheme is presented in [37]. The authors have used a patch-wise

training strategy that contains different blurred versions of input images to acquire the decision maps. By using the decision maps and the source images, the fused image is constructed. The major disadvantage of this approach is that it can only be used for the fusion of multi-focused images.

An efficient CNN-based image fusion technique is given by Prabhakar *et al.* [38], where the encoding network is used to get two feature map sequences, and a fused feature map is obtained by using the addition strategy. Herein, the resulting fused image is generated by utilizing a decoding network that comprises three CNN layers. Here, the deep network information may not have been explored completely. A densefuse network proposed by Li *et al.* [39] where the network contains a fusion layer, encoder, and decoder to get the fused images. However, this network is not able to obtain deep features. Li *et al.* [40] developed an approach to obtain the deep multi-level features which are used to fuse images. This fusion technique uses the middle layer information for the fusion scheme, where significant information is lost during the extraction of features. An end-to-end residual fusion architecture is proposed by Li *et al.* [41] where an encoder network is used to obtain the deep features from the sources at a multi-scale, and the decoder network is developed to construct the fused images. However, such an approach is unable to transfer sufficient details from the sources into the fused images. Also, an image fusion technique dependent on ResNet-152 in an NSCT domain is proposed by Gao *et al.* [42] where images from various sensors are decomposed into different frequency sub-bands. The ResNet-152 deep neural network is used to extract features in-depth for low-pass sub-bands and carry out the fusion for the same. The band-pass sub-bands are fused using the modulus maximum selection strategy. The resultant image is obtained by applying inverse NSCT on fused low-pass and band-pass sub-bands. However, this fusion scheme cannot preserve the structural information and textural details in the fused images as it uses only low-frequency sub-bands to extract deep features.

1.4.2.3 Decision-level Image Fusion Techniques

It is the process of combining the information extracted from the source images based on some decision rules to generate the fused image, which has enhanced visual contents [32]. Many techniques have been presented to report the fusion at the decision level during past decades. Rashidi *et al.* [43] proposed a multi-modal image fusion technique at the decision

level where two measures: plausibility and correctness, are jointly used to enhance the classification performance. Neagoe *et al.* [44] proposed a face recognition system based on image fusion at the decision level, which incorporates the recognition scores produced from visible channels with a thermal neural classifier. Zhao *et al.* [45] proposed a decision level image fusion technique for face recognition. The linear discriminant analysis and principal component analysis are used to obtain the face features, and the decision-level fusion rule is implemented with the source images recognition results and their confidence measures. Wang *et al.* [46] proposed a decision level image fusion technique based on the fuzzy theory where fuzzy C-means clustering is used to classify each source image and maximum membership rule is utilized to generate the fused image. However, these said techniques are unable to handle the high uncertainty in the source images. In this regard, Wang *et al.* [47] proposed a decision level image fusion technique based on the evidence theory where variance contrast, variance offset, and entropy are used as evidence. The evidence theory framework combines the evidence, and the fused image is produced according to a final decision. However, the performance of the said technique is degraded due to poorer classification results. In this regard, Vagale *et al.* [48] proposed a data fusion technique for target identification at the decision level where classification belief weights and sensor belief weights are used.

1.4.2.4 Benchmark Database for Image Fusion

In this thesis, we have considered the benchmark *TNO* dataset [2] for evaluation of the proposed scheme. The *TNO* dataset contains various challenging scenes: illumination variation, smoke, occluded objects, non-uniform lighting conditions, etc.

TNO Dataset

The *TNO* dataset consists of 63 image pairs with visual (390–700nm), near-infrared (700–100nm), and long-wave infrared (8-12μm) nighttime imagery of different surveillance and military relevant scenarios. It contains image pairs with people that are walking, running, stationary or carrying different objects, vehicles, buildings, foliage, or other artificial structures. Various sensors are used to capture the images: Athena, DHV, FEL, and TRICLOBS. These images have been registered and geometrically warped so that related image pairs have pixel-wise correspondence. Images are captured in the night-

time during various outdoor field in rural as well as urban areas.

1.4.2.5 Quantitative Measures used for Image Fusion

Although humans are the best evaluator of any vision system, it is not possible for the human being to evaluate the performance in a quantitative manner. Hence, it is necessary to evaluate methods in an objective way. Evaluation of the performance of fusion techniques is difficult as the ground truths are not always available for most challenging scenes. It is observed that in most of the SOTA techniques cited, the uses of different objective evaluation measures: Entropy (EN) [49], mutual information (MI) [49], mutual information for the discrete cosine features (FMI_{dct}) [50], mutual information for the wavelet features (FMI_w) [50], amount of artifacts added during the fusion process (N_{abf}) [51], average structural similarity index ($SSIM_a$) [52] and average edge preservation index (EPI_a) [53]. The performance of any fusion algorithm is better if the EN, MI, FMI_{dct}, $SSIM_a$, and EPI_a values are higher with lower N_{abf} value. The evaluation measures are described as follows:

Entropy The entropy measure can be described as

$$EN = -\sum_{l=0}^{L-1} P_l \log_2 P_l \ . \tag{1.2}$$

where L represents the number of grey levels and P_l is the normalized histogram of the corresponding grey level in the fused image.

Mutual Information The mutual information measure can be defined as

$$MI = MI_{I_1,F} + MI_{I_2,F} \ , \tag{1.3}$$

where I_1 and I_2 are the source images and F is the fused image.

The MI between source and fused images can be calculated as follows:

$$MI = \sum_{i,f} P_{I,F}(i,f) \log \frac{P_{I,F}(i,f)}{P_I(i) P_F(f)} \ . \tag{1.4}$$

where $P_I(i)$ and $P_F(f)$ represents the marginal probability of source image I_s and fused

image F. The joint probability of I and F represented by $P_{I,F}(i, f)$.

Mutual Information for the Discrete Cosine and Wavelet Features The FMI measure can be described as

$$FMI_F = \frac{MI(I_1; F) + MI(I_2; F)}{2}. \tag{1.5}$$

where I_1 and I_2 denote the source images. The $MI(I_s; F)$ indicates the mutual information among the source and fused images.

Amount of Artifacts Added During the Fusion Process The quantity of artifacts (N_{abf}) introduced in the fused image can be determined by considering $a = I_1$ and $b = I_2$;

$$N_{abf} = \frac{\sum_x \sum_y AM_{x,y}[(1 - \bar{Q}^{I_1F}_{x,y})\bar{w}^{I_1}_{x,y} + (1 - \bar{Q}^{I_2F}_{x,y})\bar{w}^{I_2}_{x,y}]}{\sum_{\forall x} \sum_{\forall y} (\bar{w}^{I_1}_{x,y} + \bar{w}^{I_2}_{x,y})}. \tag{1.6}$$

where $AM_{x,y} = \begin{cases} 1 & \text{if } e^F_{x,y} > e^{I_1}_{x,y} \text{ and } e^F_{x,y} > e^{I_2}_{x,y} \\ 0 & \text{otherwise} \end{cases}$. It signifies the artifacts introduced during the fusion process where the gradients of a fused image have greater strength than the input. $e^{I_1}_{x,y}$, $e^{I_2}_{x,y}$ and $e^F_{x,y}$ represents the edge strength of the visible, IR and fused images. $\bar{Q}^{I_1F}_{x,y}$ and $\bar{Q}^{I_2F}_{x,y}$ are the quantity of data that propagates from the source to fused image. The perceptual weights of the source images are interpreted by $\bar{w}^{I_1}_{x,y}$ and $\bar{w}^{I_2}_{x,y}$.

Structure Similarity Index ($SSIM$) The $SSIM$ measure can be obtained as:

$$SSIM = \sum_{i,f} \frac{2\bar{\mu}_i\bar{\mu}_f + C_1}{\bar{\mu}_i^2 + \bar{\mu}_f^2 + C_1} \cdot \frac{2\sigma_i\sigma_f + C_2}{\sigma_i^2 + \sigma_f^2 + C_2} \cdot \frac{\sigma_{if} + C_3}{\sigma_i\sigma_f + C_3}. \tag{1.7}$$

where the $SSIM$ indicates the structural similarity among the source and fused images. The standard deviation of the source and the fused images are denoted by σ_i, and σ_f, respectively. σ_{if} represents the covariance, and $\bar{\mu}_i$ and $\bar{\mu}_f$ signifies the mean values of the source and the fused image patches. The constants C_1, C_2, and C_3 are utilized to achieve the algorithm's stability. The average structural similarity index $SSIM_a$ can be

estimated as:

$$SSIM_a = \frac{SSIM(F, I_1) + SSIM(F, I_2)}{2}. \tag{1.8}$$

Edge Preservation Index The EPI quantitative metric can be obtained as:

$$EPI = \frac{\Gamma(\nabla_f - \overline{\nabla_f}, \widehat{\nabla_s} - \overline{\widehat{\nabla_s}})}{\sqrt{\Gamma(\nabla_f - \overline{\nabla_f}, \nabla_f - \overline{\nabla_f}) \cdot \Gamma(\widehat{\nabla s} - \overline{\widehat{\nabla s}}, \widehat{\nabla s} - \overline{\widehat{\nabla s}})}}, \tag{1.9}$$

$$\Gamma(F, I) = \sum_{(x,y) \in ROI} F(x, y) \cdot I_s(x, y), \tag{1.10}$$

where the high-pass filtered variant of the region of interest (ROI) acquired by 3×3 Laplacian operator in the fused and source image is indicated by ∇_f and $\widehat{\nabla_s}$. The mean of Laplacian ROI in the fused and source image is denoted by $\overline{\nabla_f}$ and $\overline{\widehat{\nabla_s}}$, respectively.

The EPI_a is the average edge preservation index and can be calculated as:

$$EPI_a = \frac{EPI(F, I_1) + EPI(F, I_2)}{2}. \tag{1.11}$$

1.4.3 State-of-the-Art Techniques for Moving Object Detection by Background Subtraction

Thermal sensors are capable of apprehending the long-wave infrared radiation reflected or emitted by the objects in the scene, which are not easy to be analyzed or detectable by a normal human vision [54]. The conventional ferroelectric barium strontium titanate (BST) thermal sensors are used in many surveillance cameras and have a good signal to noise ration (SNR) value. Thermal camera for surveillance is one of the prime areas of focus for several military and naval research. Thermal-based surveillance systems are used for two major military tasks: long-range detection of enemy vehicles and automatic target recognition (ATR). However, in recent years, the easy and cheap availability of thermal sensors and the advancement of surveillance technologies have opened up several other applications in different areas: agriculture and food industry [5], building inspection [6], gas detection [7], industrial safety [8], night surveillance [4], etc. In computer vision, the automatic surveillance from the thermal camera videos needs the background subtraction

(BGS) technique as one of the basic tools for moving object detection. For background subtraction, a sequence of image frames are initially used for modeling the background of a particular video scene. Further the target frame is compared with the constructed background model to detect the moving objects in the video. BGS techniques are greatly affected by the dynamic background condition, uncertainty in the noise level, and the multi-level brightness of the nearby pixels. Over the decades, many researchers have tried in developing robust background subtraction techniques which came a long way from the traditional approaches that used a background model specific to the video scenes. In this thesis, we have categorized the state-of-the-art techniques for the background subtraction into five different categories: parametric based [55], non-parametric based [56], sparse matrix based [57], fuzzy based [58], and deep-learning based [59].

1.4.3.1 Parametric based Background Subtraction

In parametric based techniques different set of parameters are used to establish the background model and are statistically estimated from the video frames. A mixture of Gaussian (MoG) based background modeling technique is proposed by Stauffer and Grimson [55], where the multi-valued background of a video are modeled with MoG probability density function (pdf). The parameters of the MoGs characterize the multi-valued background and are compared against the test frame's pixel value to detect the locations of the moving objects. Bhanu and Han [60] proposed an automatic human motion analysis technique for infrared image sequences. The authors have considered a modified least squares fit to estimate the 3D human walking parameters. Although the motion of the objects are analyzed but the said approach is limited by the complexity in parameter estimation. In this regard, a novel thermal video surveillance technique is studied by Xu *et al.* [61], where the objects in the thermal videos are detected using a support vector machine (SVM) followed by the normalization process and Kalman filter with mean shift is used for target tracking. Elguebaly and Bouguila [62] proposed a BGS technique for detecting the moving objects from thermal infrared videos where the finite mixtures of multidimensional asymmetric generalized Gaussian distributions are used to model the video data, and expectation–maximization (EM) algorithm is used for the parameter estimation. Subudhi *et al.* [63] proposed a robust BGS scheme where the spatio-temporal modes arising over a sequence of frames are fitted with a Gaussian *pdf* in the Wronskian

framework. A modification of the said BGS is proposed by Rout *et al.* [64], where the temporal modes are modeled with MoG and spatial modes are modeled with Wronskian function. Maddalena and Petrosino [65] proposed a local change detection technique where a self-organizing map is used to learn the background model. Further, considering the importance of parameter estimation complexity, Makantasis *et al.* [66] proposed a BGS technique where thermal responses at each pixel location are modeled using a mixture of Gaussian, and the authors have adhered to use of the Bayesian approach to estimate the parameters of the mixture structure. However, it is also true that the above-mentioned techniques are too much parameter-dependent.

1.4.3.2 Non-Parametric based Background Subtraction

The non-parametric BGS scheme is adhered to the uses of kernel or histograms-based techniques for estimating the parameters of the distribution, and hence it is found to be less complex and effective. Elgammal *et al.* [56] proposed a background subtraction technique where the parameters of the Gaussian mixture model (GMM) are estimated through kernel function. Kim *et al.* [67] proposed a background subtraction scheme where the concept of the codebook is introduced for background modeling. Even the above said approach is found to be effective in case of the non-static background, producing several misclassification results. Heikkila and Pietikainen [68] proposed the BGS scheme where the texture in image frames is modeled using local binary pattern histograms to construct the background. A contour-based BGS technique is proposed by David and Sharma [69] for foreground extraction in thermal images. The statistical BGS is used to identify the local regions-of-interest. Later on in each region, the input and background gradient information are combined to form a Contour Saliency Map, and the watershed boundaries are used to refine the contour segments. Barnich and Droogenbroeck [70] proposed a density-aware BGS technique where the spatio-contextual property is exploited with a random sampling strategy used for the background modeling. Haines and Xiang [71] proposed a robust BGS method using a Dirichlet process GMM model where the background distributions are estimated using pixel-wise non-parametric Bayesian method. It may be observed from the above analysis that, one of the main problems in surveillance is to locate the object of interest in a scene with low illumination and camouflaged foreground objects.

In this regard, St-Charles *et al.* [72] proposed a pixel level foreground segmentation scheme where spatio-temporal binary features and color information are used to detect the local changes. Subudhi *et al.* [73] proposed a statistical feature bag-based background subtraction technique where the contents of the bags are represented as average, variance, and the number of elements on the bags are used to construct the background model. Sajid and Cheung [74] proposed a background subtraction technique where multiple background models are created and the probabilities of foreground/background at each pixel location are estimated. Further, the image pixels are combined to form mega-pixels and are used to spatially denoise probability estimates to locate the object. Jiang and Lu [75] proposed a background subtraction scheme using a weight sampling mechanism. A reward-and-penalty strategy is used to reinforce active samples.

To emphasize the detection and tracking of small targets in infrared videos, Chen *et al.* [76] proposed a robust surveillance scheme where the local contrast measure and a derived kernel model is used to detect the objects in infrared videos. Further, Han *et al.* [77] proposed a small target detection technique from infrared videos, where the the difference of Gabor (DoGb) filters is proposed and improved to suppress the complex background edges, for accurate detection of objects in infrared videos. Singha and Bhowmik [78] proposed a BGS scheme, where the spatial video salient features are represented using Akin-Based Local Whitening Boolean Pattern (ALWBP), which are used to separate the foreground and background region.

1.4.3.3 Sparse Matrix based Background Subtraction

Recently, it is observed that the concept of the sparse matrix is popularly used for background construction. In a sequence of image frames, due to the motion of the moving objects, the distribution of the regions will be structurally sparse for different objects in the scene. In this regard, a local change detection technique where the spatial information in sparse outliers and the low-rank matrix is used to model the foreground and the background, respectively [57]. The concept of principal component analysis (PCA) based schemes are also devised for BGS including the binary PCA (BPCA) [79, 80]. Ebadi *et al.* [81] proposed a BGS technique, where a modified principal component analysis (PCA) uses the block sparse structure of the pixels of the moving object for local change detection. Cao et al. [82] proposed a tensor-based robust PCA for BGS for compressed

sequences, where the grouping of similar 3D patches from the background are used. Li *et al.* [83] proposed a foreground detection technique using adaptive weighted low-rank decomposition (WELD) mechanism where the gray-scale and the thermal videos given as the input to the model and the sparse outliers are separated against the modeled background to represent as moving objects. Wu and Lu [84] proposed a BGS technique where an adaptive pixel-block-based randomized arrangement is used for image sequence analysis. In the said scheme, the background model is created by an improved low-rank and block-sparse matrix decomposition.

1.4.3.4 Fuzzy based Background Subtraction

Most of the real-life image sequences are affected by illumination variation and non-static background changes which creates uncertainty in the foreground and background segmentation. Thus it may produce counterfactual inaccuracy of the BGS results. Fuzzy set theories are repute to solve the said challenges [58]. Fuzzy running average [85] for background subtraction is one of the pioneers works in state-of-the-art BGS techniques. However, the said approach fails to model the multivalued background from a sequence of image frames using fuzzy sets. In this regard, Zhang and Xu [86] proposed a fuzzy integral modeling mechanism to model the multi-valued pixels from a video scene. Further, the said approaches are found to be poor performing for non-static background conditions. Chiranjeevi and Sengupta [87] proposed a BGS technique, where the authors have used a set of fuzzy aggregated multi-feature similarity measures applied to multimodal backgrounds corresponding to the multiple models.

To deal with complex background scenes, a Type-2 fuzzy GMM model [88] is proposed to model the non-static background pixels using the color and the texture value in the video. El Baf *et al.* [89] proposed an adaptive BGS technique, where integration of Choquet integral and fuzzy set theory is used for background update. Further, Type-2 Fuzzy Gaussian Mixture Model is integrated with the Bayesian framework is proposed by Zhao et al. [90] for background subtraction. Maddalena and Petrosino [91] proposed a BGS scheme that uses spatial coherence in a self-organization map (SOM) framework. Background motion is also modeled with a fuzzy set-theoretic approach and a new histogram called fuzzy color histogram is created [92] to detect the local changes in a video scene. Although the said approach is found to be efficient but computation parameters

involved in so are difficult to estimate. Chacon-Murguia and Gonzalez-Duarte [93] proposed a neuro-fuzzy model for background subtraction where the fuzzy inference Sugeno system mimics human behavior to automatically adjust the parameters involved in the SOM detection model. Considering the complexity of the parameter estimation in neuro-fuzzy models, Zi-long et al. [94] proposed an adaptive fuzzy estimation scheme using the Takagi-Sugeno-Kang (TSK) where the parameters of the system are optimized by particle swarm optimization (PSO). Rajkumar et al. [95] proposed a BGS scheme using FCM clustering and Adaptive Network-based Fuzzy Inference System (ANFIS) classifier to deal with the multi-static background in the video scene.

Muhammet et al. [96] proposed a fuzzy BGS technique is also proposed where Choquet integral is used over a set of pixels to avoid the uncertainties in video frames. It may be observed that the color histogram of a scene can attenuate the color variations due to the non-static background which is common in a video scene. In this regard, Qiao et al. [97] proposed a background subtraction method that uses a fuzzy Color coherence vector by fuzzy c-means clustering criteria. A BGS technique is also proposed in [98], where fuzzy histograms based on fuzzy c-means clustering and the fuzzy nearness degree are used for background modeling. Recently, Subudhi et al. [99] proposed a BGS technique where an online kernelized fuzzy modal variation-based cost function is used to model the multi-valued background from a sequence of image frames.

1.4.3.5 Deep-learning based Background Subtraction

In the last few years, convolutional neural networks (CNNs) have drawn researchers' attention for local change detection [59]. The CNN-based architectures are explored to construct the background and foreground segmentation, which provides better accuracy than most of the traditional BGS techniques. Braham and Van Droogenbroeck [100] proposed a CNN-based BGS scheme where a few hand-crafted features are used for background construction. The background patches and the corresponding image patches are used to train the CNN models. A patch around each pixel of the target frame was given to the CNN that determine the label of the pixel was changed or not. Likewise, Babaee *et al.* [101] proposed a BGS scheme using deep convolutional neural network, where the background model is generated by integrating the output from SuBSENSE [72] and Flux Tensor[102] algorithm. Wang *et al.*[103] developed a foreground segmentation technique

which is based on a multi-scale CNN. It utilizes the various video frames as input with different scales, and the results were combined to predict the foreground probabilities. Further, Nguyen *et al.* [104] proposed a triplet CNN model, which is used as a motion feature extractor for local change detection. Yan *et al.* [105] proposed a BGS scheme where a deep neural network architecture is used to train the background model, and further, a cognition-based post-processing algorithm is applied to detect the moving objects. These deep learning-based BGS techniques provide better accuracy than the traditional approaches, as the deep learning framework can extract semantic and low-level features.

It may be observed that many CNN-based techniques are not an end-to-end deep learning framework and are suffered from computational complexity. In this regard, several works are proposed by researchers across the globe. Hu *et al.* [106] presented an end-to-end deep CNN for the foreground segmentation, where a 3D atrous CNN is utilized for extraction of deep features and capture the dependencies among the video frames using long short-term memory networks. Wang *et al.* [107] proposed a BGS technique where the deep and hierarchical multiscale spatial-temporal features obtained from the multiscale 3-D fully convolutional network are used to construct the background model. However, these said techniques are required many training samples to train the network. In this regard, Tezcan *et al.* [108] presented a foreground segmentation technique (BSUV-Net) where few training samples are used to train the network. A target frame and two background frames with semantic information are given input to the network for local change detection. Also, Lim and Keles [109] proposed an encoder-decoder network where an end-to-end multi-scale features learning mechanism is utilized for moving object detection.

Recently, a generative adversarial network (GAN) is also used for local change detection. Bakkay *et al.* [110] proposed conditional GAN where the generator and the discriminator are used for the foreground segmentation. However, the conditional GAN technique fails to locate the objects in continuously changing illumination in a scene. In this regard, Sakkos *et al.* [111] proposed a background subtraction technique based on triple multi-task generative adversarial network (TMT-GAN) that performs end-to-end binary classification. Also, Zheng *et al.* [112] proposed a BGS technique where the Bayesian GAN and the median filtering strategies are used to segment each pixel of the video frames as foreground/background. Parallel vision theory is further adapted to en-

hance the segmented results.

1.4.3.6 Benchmark Databases for Background Subtraction

The SOTA techniques use two challenging benchmark video databases to evaluate the performance of BGS: *changedetection.net* [113], and *Tripura University Video Dataset at Night Time (TU-VDN)* [78].

Changedetection.net Dataset

Changedetection.net [113] is a popular benchmark database used for the validation of different local change detection algorithm. This database was reported at IEEE Change Detection Workshops in 2014. This database contains (2014 dataset) with 11 different categories with each category containing almost 4 to 6 image sequences. This database contains a wide range of challenging categories including baseline, dynamic background, camera jitter, intermittent object motion, shadows, thermal, challenging weather, low frame-rate, night, PTZ, and air turbulence. It contains a total of 53 sequences with almost $\sim 1,60,000$ frames. This database contains manually annotated segmented ground truth foreground images for all the frames.

Tripura University Video Dataset at Night Time (TU-VDN)

TU-VDN [78] database contains different outdoor night videos captured by FLIR-t650sc camera of four different categories: dust, fog, rain and low light. The image sequences are captured under various atmospheric conditions such as dusty, rainy, and foggy with key challenges like flat cluttered background and dynamic background under static camera.

1.4.3.7 Quantitative Measures for Background Subtraction

Visual interpretation for evaluation is rarely found to be satisfactory to assess the quality of object detection. Such a way of evaluation lacks a quantitative measurements. Hence, it is necessary to evaluate an object/change detection method in an objective way. The measures: Precision, Recall, F-Measure, Percentage of wrong classifications (PWC),[114], Matthews correlation co-efficient (MCC) [78], and Accuracy (ACC) [78] are considered for objective evaluation of the SOTA local change detection algorithms. The performance of

any BGS scheme is better if the Precision, Recall, F-Measure, MCC, and ACC values are higher with lower PWC value. Precision is described as a fraction of the retrieved instances that are relevant, whereas Recall is defined as the fraction of relevant instances that are retrieved. F-Measure combines Precision and Recall and is the harmonic mean of Precision and Recall. The percentage of wrong classifications is defined as the ratio of instances misclassified over all the instances. Matthews correlation co-efficient is the correlation co-efficient between the predicted and true instances. Accuracy is the ratio between the accurately predicted instances and all the instances. These evaluation measures can be calculated as;

$$\text{F-Measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{1.12}$$

where $\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$ and $\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$.

$$\text{PWC} = \frac{100 \times (\text{FP} + \text{FN})}{\text{FP} + \text{FN} + \text{TP} + \text{TN}}. \tag{1.13}$$

$$\text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP} + \text{FP}) \times (\text{TP} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TN} + \text{FN})}} \tag{1.14}$$

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{1.15}$$

where true positive (TP) is the number of pixels correctly labeled as an object class, false positive (FP) is the number of pixels incorrectly classified as an object class, true negative (TN) is the number of pixels correctly labeled as background class, and false negative (FN) is the number of pixels wrongly classified as background class. The TP, FP, TN, and FN are determined by comparing the ground-truth images provided in the datasets with the output images obtained by different techniques considered for the evaluation.

1.5 Scope of The Thesis

This thesis has a total of four contributing chapters are as follows:

1.5.1 Contrast Preservation with Intensity Variation approach for Pixel Level Image Fusion

In this work, we have proposed two pixel level fusion schemes: fuzzy edge preserving intensity variation approach and weighted combination of maximum and minimum value selection strategy. In the proposed fuzzy edge preserving intensity variation approach, the salient feature map is obtained by analyzing the spatial inter-dependency between the visible and IR images. However, the salient feature map is unable to retain the edges from the source images. Therefore, we have explored the concept of fuzzy edge on visible image to obtain its edges. The fused image is generated by combining the salient feature map and edges of the visible image. Again in the weighted combination of maximum and minimum value selection strategy, initially, a center sliding window is used in the IR image. Then the moving block-based average operator is used for each center sliding window. The detail feature map is obtained by using the maximum selection strategy between the output of the block-based average operator with the corresponding location of the visible image. The detail feature map is not able to preserve the subtle details from the source images. So the intermediate feature map is acquired by applying the minimum selection strategy among the source images. Eventually, the fused image is obtained by using the weighted-average fusion approach among the detailed and intermediate feature maps.

The proposed schemes are evaluated with challenging source pairs available at the benchmark *TNO* database [2]. The efficacy of the proposed fuzzy edge preserving intensity variation approach is validated against eight state-of-the-art schemes, and the efficiency of the proposed weighted combination of maximum and minimum value selection strategy is corroborated against the seven existing SOTA techniques. The performance of the proposed techniques is validated qualitatively and quantitatively in order to justify our findings.

1.5.2 Integration of Multi-scale Features with Deep Learning Architecture for Feature Level Image Fusion

In this work, we have proposed two feature level fusion schemes: integration of bi-dimensional empirical mode decomposition with two streams VGG-16 and non-subsampled contourlet transform induced two streams ResNet-50 network. In the proposed integration of bi-dimensional empirical mode decomposition with two streams VGG-16 scheme, we have adhered to the bi-dimensional empirical mode decomposition (BEMD) scheme to decompose the source images into several intrinsic mode functions ($IMFs$) at different frequency bands. The proposed BEMD strategy with VGG-16 architecture explores features in-depth on frequency domain at various levels and can handle the high uncertainty in the source images. The proposed deep multi-level fusion strategy constructs the weight maps to preserve the correlative data accurately from the images of different sensors and provides a detailed fusion map. The minimum selection strategy among these detailed maps retains the standard information and reduces the superfluous data. Again, in the proposed non-subsampled contourlet transform induced two streams ResNet-50 network algorithm, the source images corresponding to the visual and thermal sensors are decomposed into the multi-directional, multi-scale, and shift-invariant coefficients using the non-subsampled contourlet transform (NSCT). The low frequency and high-frequency coefficients of the NSCT are fed into a two-stream ResNet-50 architecture to extract multi-layer deep features. The use of NSCT followed by the deep neural architecture on frequency domain led to the extraction of deep features in multi-direction with a different scale which is one of the essential requirements for the image fusion scheme. In the proposed technique, the considered ResNet-50 network consists of five convolutional blocks. At each block of the ResNet-50 network, the sum of the absolute difference (SAD) and the moving block-based average (MBBA) operator are used to retain the details of the source images. A weight map construction strategy has been proposed, where the normalization operator and the bicubic interpolation are utilized to capture the complementary information present in the source images. The feature maps are thus obtained by using the weight maps and the source images. Eventually, the maximum and the minimum selection strategy among these feature maps are used over all the convolutional blocks of the ResNet-50 to generate the fused image. Both the proposed techniques provide a fused image with lesser artifacts for a pair of input IR and visible images.

Different experiments were carried out on the *TNO* benchmark database [2] to estimate the efficacy of the proposed algorithms. The efficiency of the proposed integration of bi-dimensional empirical mode decomposition with two streams VGG-16 technique is corroborated against fifteen existing state-of-the-art fusion techniques, and the efficacy of the proposed non-subsampled contourlet transform induced two streams ResNet-50 network algorithm is validated against the existing ten state-of-the-art techniques. The competency of the proposed algorithm is estimated using qualitative and quantitative assessments and found to be efficient.

1.5.3 Kernel Induced Possibilistic Fuzzy Associate Background Subtraction for Moving Object Detection

In this work, we have proposed a kernel induced possibilistic fuzzy associate background subtraction for video scene unsupervised background subtraction technique to detect the local changes in fixed camera captured sequences. The proposed scheme follows two stages: background training and foreground segmentation. In the background construction stage, each pixel is modeled using a possibilistic fuzzy cost function in kernel induced space. The use of induced kernel function projects the low dimensional data into a higher dimensional feature space and the use of possibilistic function will construct a robust background model based on the density of the data in temporal direction avoiding the noisy and outlier points. The performance of the proposed scheme is tested on the benchmark database: *changedetection.net* [114]. The effectiveness of the proposed scheme is evaluated on different performance evaluation measures. We corroborate our findings by comparing them against twenty-nine existing state-of-the-art BGS techniques.

1.5.4 Multi-Scale Deep Learning Architecture based Background Subtraction for Moving Object Detection

In this work, we have proposed two multi-scale deep learning architectures based background subtraction for moving object detection: modified ResNet-152 network with hybrid pyramidal pooling and multi-scale contrast preserving deep learning architecture. In the proposed modified ResNet-152 network with hybrid pyramidal pooling algorithm, we have designed an encoder-decoder type deep learning architecture with transfer learning

for background subtraction (BGS). Here, a modified ResNet-152 network is considered as an encoder to enhance the use of high-frequency components for the foreground segmentation. We have developed a multi-scale features extraction (MFE) mechanism block, a hybridization of pyramidal pooling architecture (PPA), and various atrous convolutional layers to extract features at various scales. The use of PPA enhances the performance of an MFE block which can handle various challenging scenes effectively. We have also proposed an efficient decoder consisting of stacked transposed convolution layers (Tconvs) to project from feature-level to pixel-level, predicting a score map. Then, a threshold is applied on the score map to get the binary class labels as the background and foreground. The higher blocks' extracted features have semantic information but lack of providing low-level features that are generally important for the foreground segmentation. Therefore, shortcut connections followed by global average pooling (GAP) drive the low-level features from the encoder network to the decoder network.

Again, in the proposed multi-scale contrast preserving deep learning architecture, we have designed an encoder network that considers a hybrid of convolution and atrous convolution blocks to preserve both sparse and dense features of a video with skip connection. The proposed encoder with the multi-scale contrast preservation block is able to keep multi-scale contrast features with less training loss. Further, the proposed decoder network accurately projects the extracted features at different layers into pixel-level. The proposed end-to-end model efficiently provides a binary change detection map.

The proposed techniques are corroborated by testing it on two benchmark databases: *changedetection.net* [113], and *Tripura University Video Dataset at Night Time (TU-VDN)* [78]. The effectiveness of the proposed modified ResNet-152 network with hybrid pyramidal pooling algorithm is evaluated by comparing the results obtained by it with thirty-one existing state-of-the-art techniques, and the efficacy of the proposed multi-scale contrast preserving deep learning architecture is determined by comparing it with twenty-eight state-of-the-art BGS techniques. The performance of the proposed techniques is validated qualitatively and quantitatively, and found to be efficient.

1.6 Organization of the Thesis

The rest of the thesis organization is as follows. Chapter 2 describes the proposed contrast preservation with intensity variation approaches for pixel level image fusion: fuzzy edge preserving intensity variation approach and weighted combination of maximum and minimum value selection strategy. The proposed integration of multi-scale features with deep learning architectures for feature level image fusion: integration of bi-dimensional empirical mode decomposition with two streams VGG-16 and non-subsampled contourlet transform induced two streams ResNet-50 network are discussed in Chapter 3. Chapter 4 illustrates the proposed kernel induced possibilistic fuzzy associate background subtraction for moving object detection. The proposed multi-scale deep learning architectures based background subtraction for moving object detection: modified ResNet-152 network with hybrid pyramidal pooling and multi-scale contrast preserving deep learning architecture are presented in Chapter 5. Chapter 6 draws the overall conclusions and scope for future work.

Chapter 2

Contrast Preservation with Intensity Variation approach for Pixel Level Image Fusion

2.1 Introduction

A brief introduction of the existing pixel level fusion techniques are discussed in Chapter 1. It can be summarized from the state-of-the-art pixel level fusion techniques that they are unable to retain sufficient edge details in the fused images. The pixel level fusion techniques add more artifacts in the fused images during the fusion process. Also, the said techniques are found to be producing low contrast fused images. Such fused image are unable to be used for many higher order applications: object detection, object tracking, object recognition, etc. One of such examples are provided in Figure 2.1. Figure 2.1 (a) and (b) depict a pair of visible and IR images considered for experimentation. The fused image obtained by ratio of low-pass pyramid (RP) [24] (an existing pixel level fusion technique) and corresponding histogram are presented in Figure 2.1 (c) and (d). It may be observed from Figure 2.1 (c) and (d) that, the fused image produced by the said technique is incapable of preserving the edge details with low contrast. A zoomed portion of the fused image depict that, the fused image does not carry meaningful information and is also not suitable for human visible perception.

In this context, we have proposed two pixel level fusion schemes: fuzzy edge preserv-

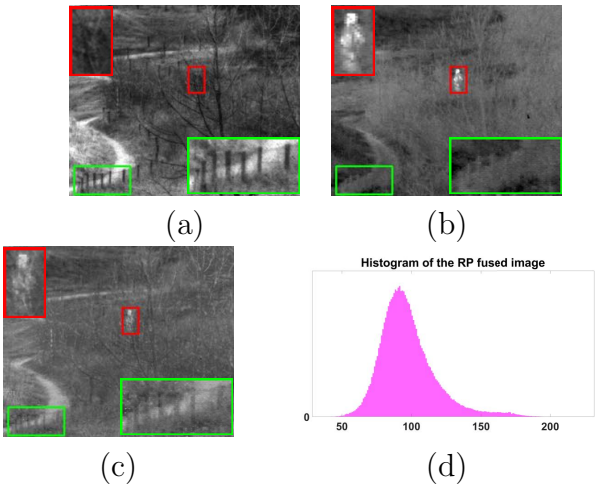


Figure 2.1: Visual analysis of (a) Visible image, (b) IR image (c) Fused image obtained by the RP technique, and (d) Histogram of the fused image.

ing intensity variation approach and weighted combination of maximum and minimum value selection strategy, to preserve and retain the edge details in the fused image. In the proposed fuzzy edge preserving intensity variation approach, the maximum selection strategy is used among the median of the infrared image with the corresponding pixel on visible image to generate the salient feature map (SFM). The edges of the visible image is obtained in the next stage of the algorithm using a Fuzzy edge technique. Finally the fused image is obtained by combining the salient feature map and the edges of the visible image. In the weighted combination of maximum and minimum value selection strategy, the detail feature map is obtained by utilizing the maximum selection strategy between the pixel-intensity of visible image and block-based average of center sliding window. In the next stage, the source images are compared by using the minimum selection strategy to obtain the intermediate feature map. Eventually, the fused image is achieved by utilizing the weighted-average technique among the detail and intermediate feature maps.

The proposed schemes are tested on the benchmark *TNO* database. The effectiveness of the proposed fuzzy edge preserving intensity variation approach is validated against eight state-of-the-art schemes, and the efficiency of the proposed weighted combination of maximum and minimum value selection strategy is corroborated against the seven existing techniques. The performance of the proposed techniques is validated qualitatively and quantitatively in order to justify our findings. It is observed that the proposed algorithms are attained higher accuracy against the considered state-of-the-art techniques.

The rest of this chapter is organized as follows. The proposed pixel level fusion algo-

rithms are discussed in Section 2.2. Section 2.3 presents the results and discussions with future works. The conclusion of the proposed works is carried out in Section 2.4.

2.2 Proposed Contrast Preservation with Intensity Variation approach for Pixel Level Image Fusion

In this chapter, we have proposed two pixel level fusion schemes: fuzzy edge preserving intensity variation approach and weighted combination of maximum and minimum value selection strategy, to preserve and retain the edge details in the fused image. Here we assumed that, the I_s, s denotes the source images from various sensors, where $s \in \{1, 2\}$, 1 corresponds to the visible image, 2 corresponds to the IR image, and F is the fused image.

2.2.1 Proposed Fuzzy Edge Preserving Intensity Variation Approach

In this chapter, we propose a novel pixel level fusion method whose graphical illustration is presented in Figure 2.2. The detailed description of different stages of the proposed scheme are narrated as follow.

2.2.1.1 Spatial Domain Analysis

In the initial stage of the proposed scheme, we considered a center sliding window in IR image $I_2(x, y)$ of size $w \times w$. To include the border pixels in the operation we initially zero padded the IR image. The number of rows and columns to be zero-padded on each side of the IR image based on the size of the center sliding window as $\frac{w-1}{2}$.

The spatial pixel distribution of salient feature map $SFM(x, y)$ can be obtained using a maximum selection strategy between the median of the center sliding window with the corresponding position in the visible image as,

$$SFM(x,y) = \begin{cases} I_1(x,y) & \text{if } I_1(x,y) > \underset{(x,y)\in(w\times w)}{\text{median}}\, I_2(x,y) \\ \underset{(x,y)\in(w\times w)}{\text{median}}\,(I_2(x,y)) & \text{otherwise} \end{cases}. \tag{2.1}$$

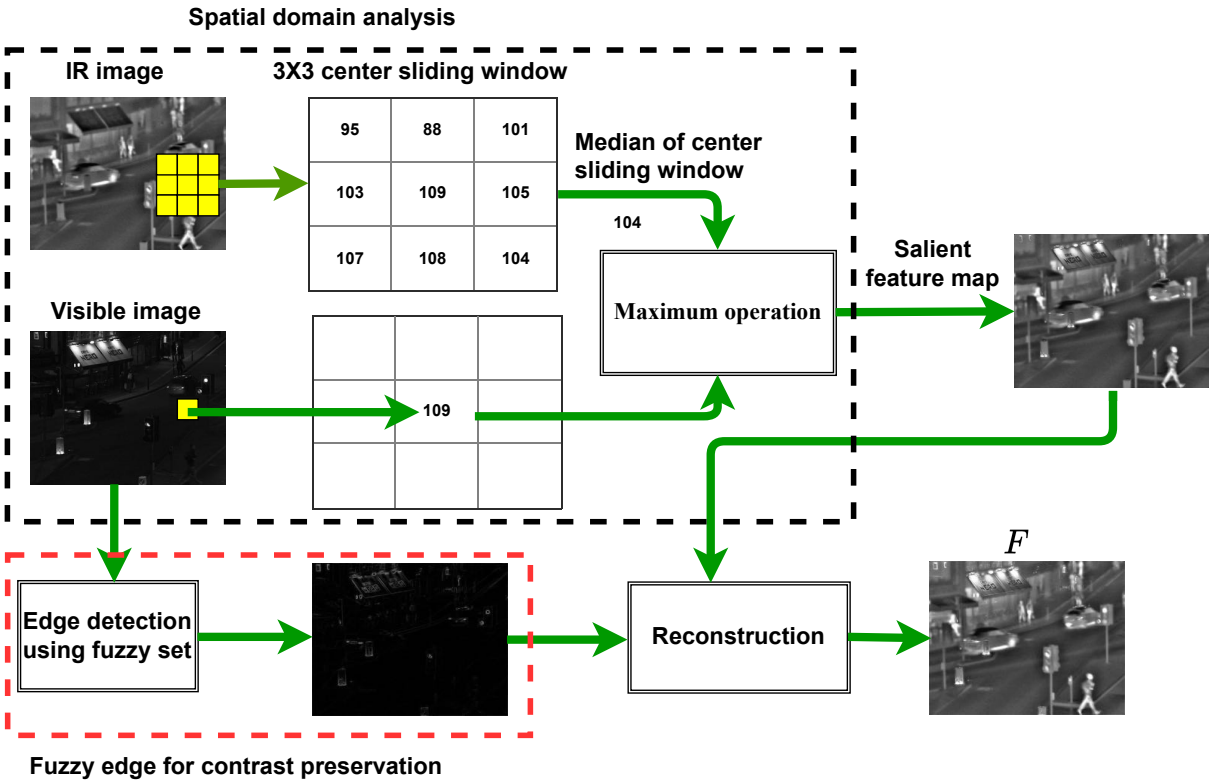


Figure 2.2: Block diagram of the proposed fuzzy edge preserving intensity variation approach.

The salient feature map obtained in this stage is unable to retain the edges from the source images. Therefore, in the next stage, we have explored the concept of fuzzy edge to retain the contrast of the images.

2.2.1.2 Fuzzy Edge for Contrast Preservation

The images captured from multiple sensors possess uncertainty within a pixel due to the multi-valued brightness level. It is obvious that a deterministic method of obtaining edges may not give a better results in image fusion. Hence, it is required to explore the capabilities of fuzzy sets theoretic approaches. In the proposed algorithm, we have used a fuzzy edge preservation mechanism to obtain the edge details of the visible image.

Let us assume fuzzy set B of the universe Y can be represented as

$$B = \{(\mu_B(y), y), \forall \in Y\}, \tag{2.2}$$

where the characteristic function $\mu_B(y)$, $(0 \leq \mu_B(y) \leq 1)$ in fact, can be viewed as a

weighting coefficient that reflects the ambiguity in a set, and as it approaches unity, the grade of membership of an event in B becomes higher. For example, $\mu_B(y) = 1$ indicates a strict containment of the event y in B. If on the other hand, y does not belongs to B, $\mu_B(y) = 0$. Any intermediate value would represent the degree to which y could be a member of B. If $\mu_B = 0.5$, then y is called a crossover point. Fuzzy set theory can be extended to images where an image I of size $M \times N$ with the maximum grey level L can be treated as a two-dimensional array of fuzzy singletons. Each fuzzy singleton with a membership function representing the degree of having illumination level l, where $l = 0$, $1, \cdots L-1$. Therefore, an image can be represented as

$$I = \bigcup_m \bigcup_n p_{mn}/y_{mn} \ , \tag{2.3}$$

$$m = 0,\, 1, \cdots M-1; \quad n = 0,\, 1, \cdots N-1.$$

where p_{mn}/y_{mn} denotes the grade of owning some property p_{mn} by the $(m,n)^{th}$ pixel intensity y_{mn}. To derive the property p_{mn}, we have used similar kind of expression used in [115] and the expression for p_{mn} can be given as;

$$p_{mn} = \left[1 + \frac{|I_1(x,y) - \min\limits_{(x,y)\in(w\times w)} I_1(x,y)|}{F_d}\right]^{-F_e}, \tag{2.4}$$

The property plane is determined using max or min operator. F_e indicates the exponential and F_d signifies the denominational fuzzifiers, respectively. By varying the value of fuzzifiers we can control the fuzziness in the property plane. The positive constant F_e and F_d are independent of the pixel locations. From (2.4), it is illustrious, for $\min\limits_{(x,y)\in(w\times w)} I_1(x,y) = 0$, p_{mn} equal to α is finite positive quantity. Hence, the range of p_{mn} is $(\alpha,\, 1)$ instead of $(0,\, 1)$. After getting the property plane, the edge $E(x,y)$ of an image can be determined using the operation;

$$E(x,y) = (L-1)\left\{1 - \left[1 + \frac{|I_1(x,y) - \min\limits_{(x,y)\in(w\times w)} I_1(x,y)|}{F_d}\right]^{-F_e}\right\} \tag{2.5}$$

$$L = 2^n \ . \tag{2.6}$$

Here L is the maximum grey level of the image and n is the number of bits required to represent each pixel.

2.2.1.3 Fused Image Generation

In the final stage the fused image $F(x, y)$ can be constructed by combining $SFM(x, y)$ obtained from (2.1) and the $E(x, y)$ obtained from (2.5), using,

$$F(x, y) = A[SFM(x, y) + E(x, y)]. \tag{2.7}$$

where $A = \frac{L-1}{2L-2}$. A is used to make $F(x, y)$ in the range (0, 255).

2.2.2 Proposed Weighted Combination of Maximum and Minimum Value Selection Strategy

In this chapter, we also propose a novel pixel level fusion method whose graphical illustration is presented in 2.3. The subsequent steps of the proposed algorithm are described as follow.

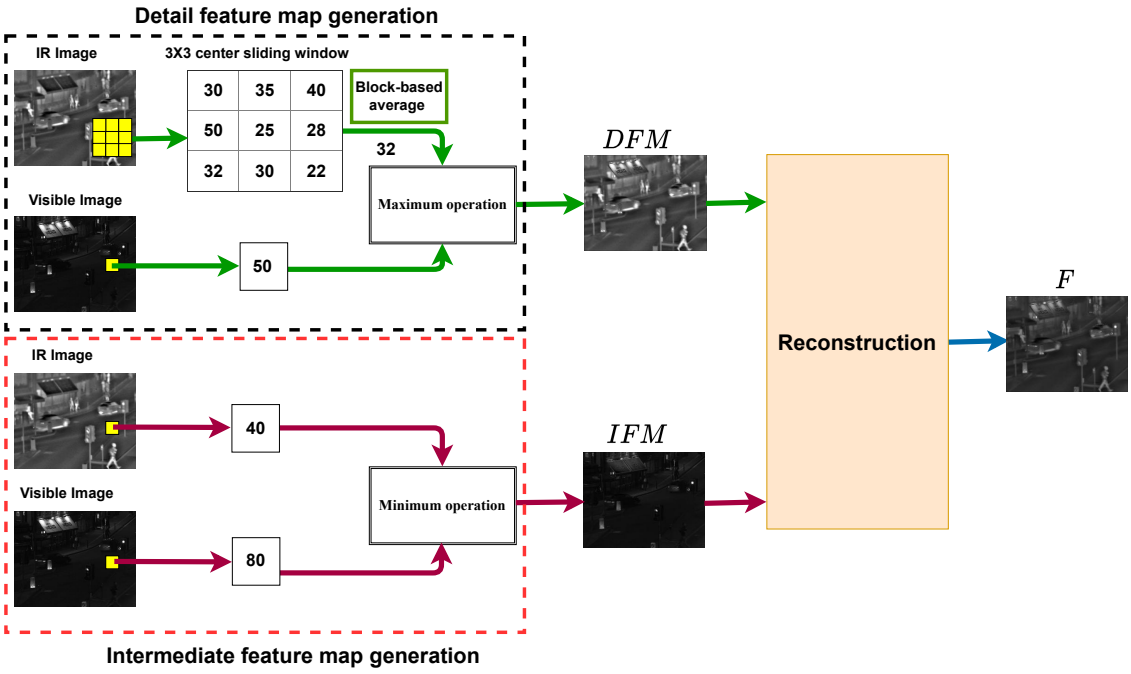


Figure 2.3: Block diagram of the proposed weighted combination of maximum and minimum value selection strategy.

2.2.2.1 Detail Feature Map Generation

The thermal radiation information normally represented by the pixel values. Hence, the objects are clearly identified in the IR image because of the grey level variation among the background and the objects. This inspired us to propose a fusion scheme where the fused image which have the comparable pixel values dissemination with the provided IR image. In the proposed scheme, we consider a center sliding window $w \times w$ in the IR image. To include border pixels in operation, we initially zero-padded the IR image. The rows and columns are zero-padded with $\frac{w-1}{2}$. We utilize the block-based average operator ($BBAO$) in each center sliding window. The $BBAO$ obtained at each center sliding window is calculated as,

$$BBAO(x,y) = \frac{\sum_{p=-\frac{w-1}{2}}^{\frac{w-1}{2}} \sum_{q=-\frac{w-1}{2}}^{\frac{w-1}{2}} I_2(x+p, y+q)}{w^2}. \tag{2.8}$$

where I_2 indicates the IR image, and (x, y) indicates the pixel location.

Then the detail feature map DFM is obtained by applying the maximum selection strategy between the $BBAO$ and pixel value at any pixel location (x, y) in the visible image as,

$$DFM(x,y) = \begin{cases} BBAO(x,y) & \text{if } BBAO(x,y) > I_1(x,y) \\ I_1(x,y) & \text{otherwise} \end{cases}, \tag{2.9}$$

where $BBAO(x, y)$, $I_1(x, y)$ represent the block-based average operator and pixel value in the visible image at location (x, y), respectively. The salient features that we have obtained in the above process are not suitable for human visible perception, which is shown in Figure 2.3. So in the next stage, we need to preserve essential details from the source images, which may be more informative.

2.2.2.2 Intermediate Feature Map Generation

To enhance visual contents in the fused image, the background information is the important requirement. In great detail the background is characterized by the visible image as compared to the IR image. Therefore, to get the detailed information of the background, In this stage, the intermediate feature map IFM is generated by using the minimum

selection strategy among the source images and can be given as,

$$IFM(x,y) = \begin{cases} I_2(x,y) & \text{if } I_2(x,y) < I_1(x,y) \\ I_1(x,y) & \text{otherwise} \end{cases}, \tag{2.10}$$

The information contents that we obtained in the IFM have complementary properties as compared to the information contents of DFM. Both DFM and IFM contains the standard features and the redundant information as well. So to reconstruct the fused image, we use the weighted-average technique, which is discussed below.

2.2.2.3 Fused Image Generation

The salient features, such as DFM and IFM, we obtained from the source images, contains the standard features. Here, we consider the weighted-average technique to fuse the salient features for reconstructing the fused image and can be given as;

$$F(x,y) = \beta_1 DFM(x,y) + \beta_2 IFM(x,y), \tag{2.11}$$

where β_1 and β_2 denote the weight values for the detail feature map and intermediate feature map. To retain the common information and minimize the redundant information, in this article, we considered $\beta_1 = 0.5$ and β_2=0.5.

2.3 Results and Discussions

In this section, we test the performance of the proposed schemes on a publicly accessible *TNO* database collected from [2]. To test the efficiency of the proposed fuzzy edge preserving intensity variation approach, we compared the results obtained by it with those of eight existing state-of-the-art fusion techniques: Laplacian pyramid (LP) [23], filter-subtract-decimate pyramid (FSD) [26], gradient pyramid (GP) [26], ratio of low-pass pyramid (RP) [24], contrast pyramid (CP) [25], morphological pyramid (MP) [26], discrete wavelet transform (DWT) [27], and shift invariant discrete wavelet transform (SI-DWT) [27]. To justify the efficiency of the proposed weighted combination of maximum and minimum value selection strategy, we compared it against seven state-of-the-art image fusion techniques: Cross bilateral filter (CBF) [116], Weighted least square (WLS)

[117], convolutional sparse representation (CSR) [34], ratio of low-pass pyramid (RP) [24], ratio of low-pass pyramid-sparse representation (RP-SR) [24], Convolutional sparsity based morphological component analysis (CS-MCA) [35] and Convolutional neural network (CNN) [36].

All the experiments in this chapter are performed on a PC with 3.20 GHz Intel Core CPU and 16 GB RAM. To measure the performance of our proposed algorithm against the considered existing state-of-the-art techniques, we have performed both subjective and objective evaluation. With the goal of quantitative comparison between the proposed techniques and the existing fusion strategies, various evaluation measures: Entropy (EN) [49], mutual information (MI) [49], mutual information for the discrete cosine features (FMI_{dct}) [50], mutual information for the wavelet features (FMI_w) [50], amount of artifacts added during the fusion process (N_{abf}) [51], average structural similarity index ($SSIM_a$) [52] and average edge preservation index (EPI_a) [53] are used. The fusion performance is better when the numerical values of evaluation measures such as: EN, MI, FMI_{dct}, FMI_w, $SSIM_a$, EPI_a are higher with lower value of N_{abf}.

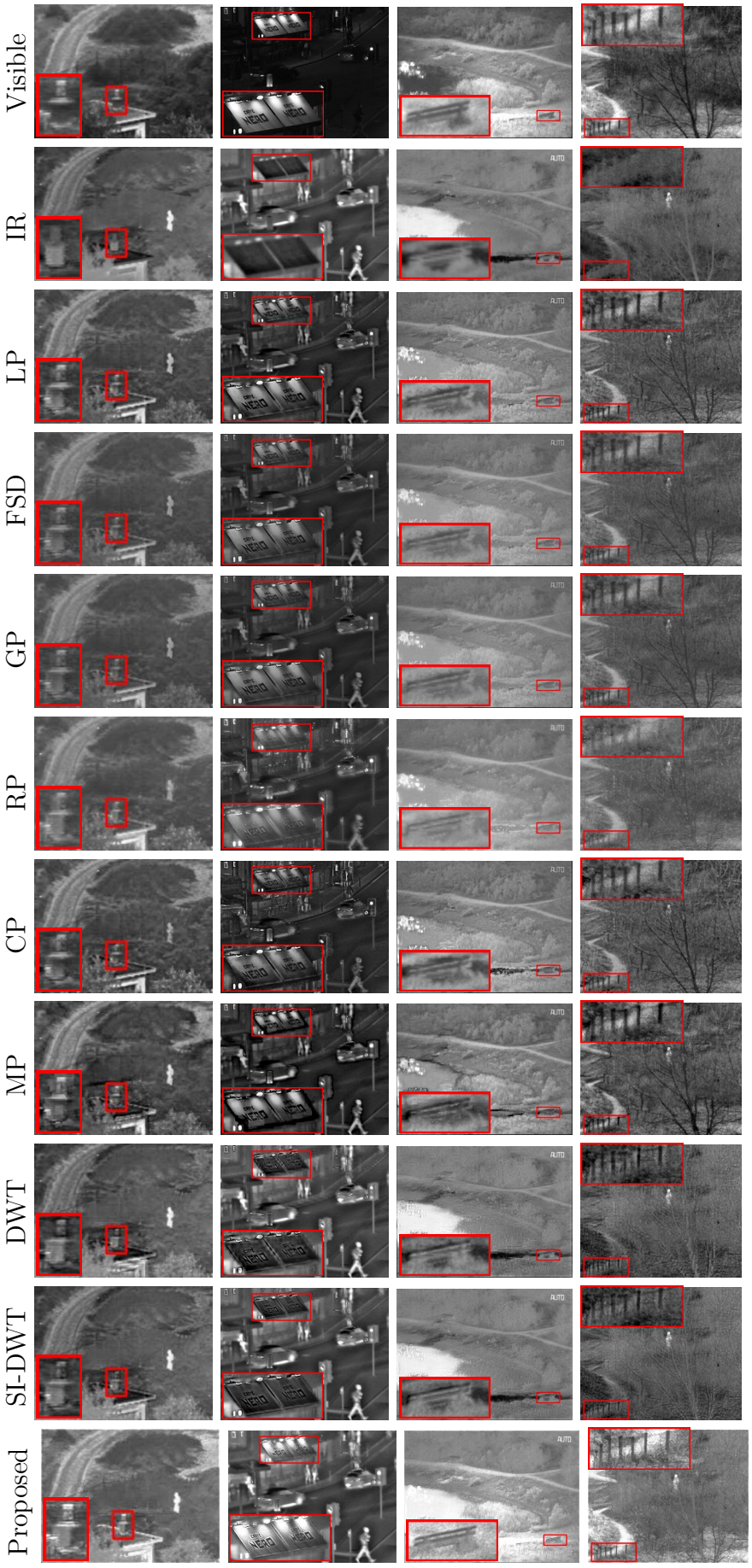


Figure 2.4: Visual analysis of results on the Nato_camp, Street, Lake, and Sandpath images (from left to right). From top to bottom: Visible images, IR images, fused images obtained by LP, FSD, GP, RP, CP, MP, DWT, SI-DWT, and the proposed fuzzy edge preserving intensity variation approach.

2.3.1 Qualitative illustration of Fuzzy Edge Preserving Intensity Variation Approach

The source images and corresponding fused images obtained by the state-of-the-art techniques and the proposed approach are shown in Figure 2.4. The first and second rows are the visible and IR images of Nato_camp, Street, Lake, and Sandpath images taken from the *TNO* database used for fusion. It can be seen from Figure 2.4 that, the foreground objects are not clearly highlighted in the fused images obtained by the LP, FSD, and GP techniques. The fused images acquired by the RP, CP, and MP techniques contain more artifacts, and the salient features are not clear. The fused images achieved by the DWT and SI-DWT techniques found to have blurred details with more noise. However, the fused images obtained by the proposed scheme precisely preserve the foreground and background information with reduced artifacts against the existing state-of-the-art fusion schemes.

Histogram analysis is an apt strategy to decide the contrast of any image. Therefore, the same has been used herein. Figure 2.5 represents the histogram plot of the eight existing state-of-the-art and proposed techniques for the Kaptein 1123 image. From this Figure, it may be observed that, the proposed algorithm produces high contrast fused image as compared to the existing state-of-the-art fusion techniques.

2.3.2 Quantitative comparison of Fuzzy Edge Preserving Intensity Variation Approach

The quantitative evaluation of the proposed fuzzy edge preserving intensity variation approach and the state-of-the-art fusion techniques are achieved using four evaluation measures: EN, MI, FMI_{dct} and FMI_{w}.

The values of EN, MI, average of FMI_{dct} and average of FMI_{w}, are reported in Table 2.1 - 2.4 as obtained using eight state-of-the-art techniques and the proposed technique, where best values are indicated in bold. Also, the graphical comparison on EN and MI measures among different techniques is shown in Figure 2.6. As we can see from Table 2.1-2.2 and Figure 2.6, the proposed technique has comparably a good ENs and MIs than other state-of-the-art techniques but smaller value of EN compared to the RP method for tank image only. From Table 2.3-2.4, it may be found that the average values of FMI_{dct}

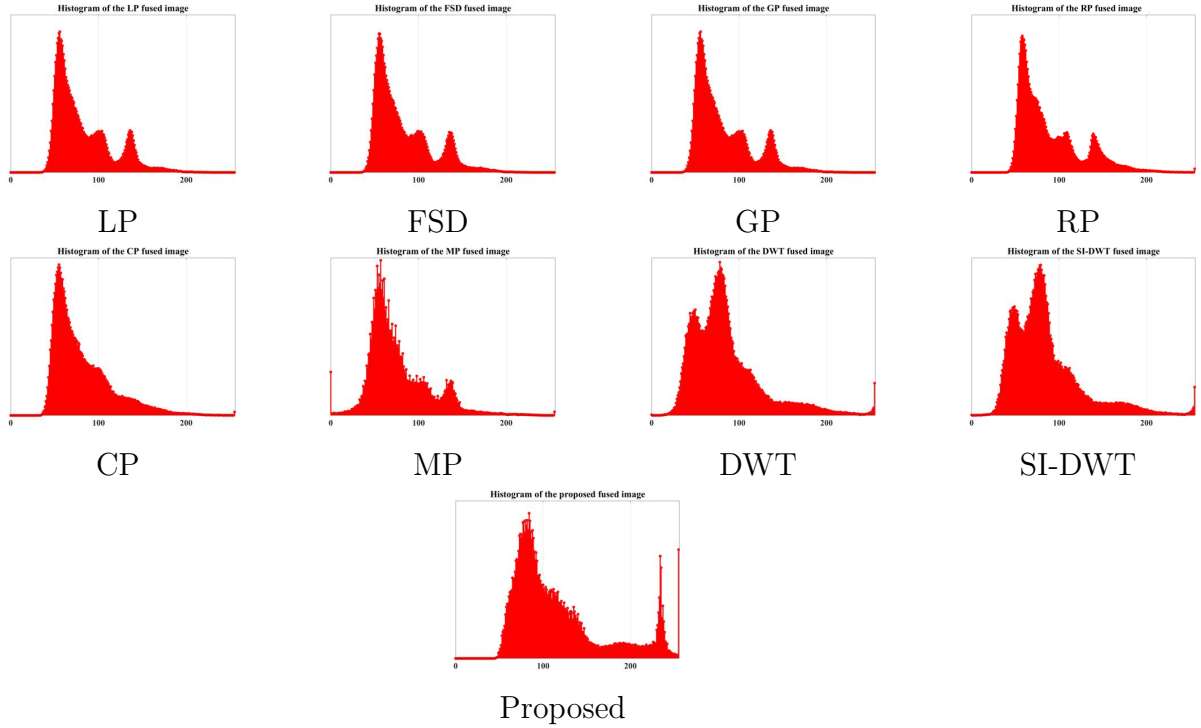


Figure 2.5: The histogram of Kaptein 1123 fused images obtained by (a) LP, (b) FSD, (c) GP, (d) RP, (e) CP, (f) MP, (g) DWT, (h) SI-DWT, and (i) proposed fuzzy edge preserving intensity variation approach.

and FMI_w obtained for our proposed method are better as compared to the considered state-of-the-art methods. From the quantitative comparison, our proposed technique has better fusion performance than the state-of-the-art techniques.

Table 2.1: Quantitative comparisons of entropy

Algorithm	Kaptein 1123	Nato_camp	Street	Lake	Sandpath	Duine	Tank
LP	6.731	6.571	6.249	6.680	6.488	5.954	7.436
RP	6.720	6.442	6.161	6.636	6.231	5.768	**7.461**
CP	6.696	6.631	6.110	6.772	6.556	5.949	7.334
FSD	6.603	6.315	5.978	6.579	6.218	5.776	7.392
GP	6.598	6.307	5.975	6.579	6.211	5.772	7.387
MP	6.826	6.845	6.449	6.847	6.710	6.129	7.267
DWT	7.036	6.843	6.764	6.667	6.588	5.961	6.283
SI-DWT	7.014	6.813	6.745	6.646	6.540	5.945	6.359
Proposed	**7.170**	**6.910**	**6.843**	**6.934**	**6.805**	**6.157**	7.448

Table 2.2: Quantitative comparisons of mutual information

Algorithm	Kaptein 1123	Nato_camp	Street	Lake	Sandpath	Duine	Tank
LP	1.781	1.395	1.906	1.601	0.955	1.297	1.881
RP	1.886	1.433	1.883	1.723	0.971	1.386	1.667
CP	1.442	1.337	1.369	1.630	0.941	1.288	1.392
FSD	1.938	1.432	2.234	1.877	1.039	1.344	2.205
GP	1.956	1.445	2.248	1.892	1.055	1.358	2.261
MP	1.627	1.342	1.780	1.285	0.814	1.216	1.945
Proposed	**3.199**	**2.045**	**3.324**	**2.820**	**1.725**	**1.892**	**2.519**

Table 2.3: Quantitative comparisons of mutual information for the discrete cosine features

Images	LP	RP	CP	FSD	MP	Proposed
Kaptein 1123	0.291	0.232	0.247	0.308	0.231	0.308
Nato_camp	0.279	0.241	0.268	0.297	0.194	0.286
Sandpath	0.269	0.224	0.245	0.285	0.212	0.305
Duine	0.304	0.263	0.302	0.318	0.235	0.309
Tank	0.179	0.152	0.145	0.192	0.154	0.232
Average	0.264	0.222	0.241	0.280	0.205	**0.288**

Table 2.4: Quantitative comparisons of mutual information for the wavelet features

Images	LP	RP	CP	FSD	MP	DWT	SI-DWT	Proposed
Sandpath	0.327	0.293	0.344	0.336	0.275	0.340	0.367	0.345
Tank	0.309	0.260	0.302	0.322	0.280	0.258	0.279	0.325
Average	0.318	0.277	0.323	0.329	0.278	0.299	0.323	**0.335**

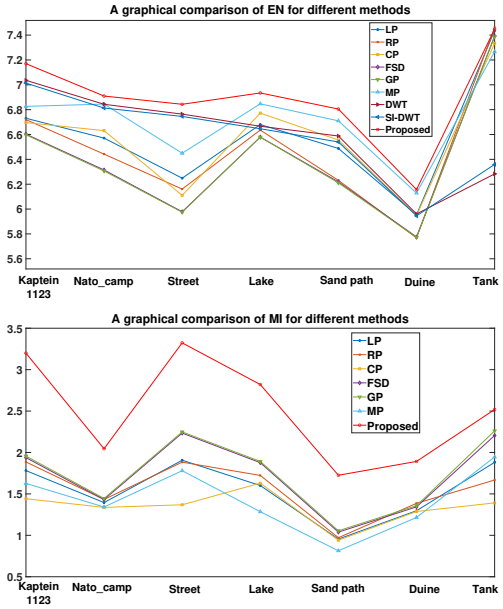


Figure 2.6: Quantitativ comparisons of EN and MI on Kaptein 1123, Nato_camp, Street image, Lake, Sandpath, Duine and Tank.

Figure 2.7: Visual analysis of results on the Kaptein 1123, Marne and Bench images (from left to right). From top to bottom: Visible images, IR images, fused images obtained by CBF, RP, RP-SR, CS-MCA, and proposed weighted combination of maximum and minimum value selection strategy.

2.3.3 Qualitative illustration of Weighted Combination of Maximum and Minimum Value Selection Strategy

The source images and the results obtained by the state-of-the-art techniques as well as the proposed algorithm are presented in Figure 2.7. The first and second rows are the visible and IR images used for fusion. As we can perceive from Figure 2.7 that the fused images obtained by the CBF method cannot retain the edge details from the source images and introduce more artifacts during the fusion process. The fused images obtained by the RP technique preserve ambiguous visual content, which are not suitable for human visual perception. The results acquired by the RP-SR scheme have many isolated points which degraded the quality of background and foreground information. The visual impression attained by the CS-MCA technique is not clear because of the ringing artifacts. However, the resultant images obtained by the proposed scheme found to have enhanced features with reduced artifacts against the existing fusion schemes. Also, the fused images obtained by the proposed scheme strongly correlate with source images and look more natural compared to the state-of-the-art techniques.

2.3.4 Quantitative comparison of Weighted Combination of Maximum and Minimum Value Selection Strategy

We have considered four evaluation measures: FMI_{dct} [50], N_{abf} [51], average structural similarity index $SSIM_a$ [52], and EPI_a [53] to evaluate the performance of the proposed technique. The average values of FMI_{dct}, N_{abf}, $SSIM_a$, and EPI_a obtained by the fused image resulting from the state-of-the-art and proposed weighted combination of maximum and minimum value selection strategy are reported in Table 2.5 where the finest values of evaluation measures are represented in bold. From Table 2.5 we can perceive that the FMI_{dct}, N_{abf} and $SSIM_a$ of the proposed algorithm have the best average values as compared to the seven state-of-the-art techniques. But, the average EPI_a value of our algorithm is higher than all the existing techniques excluding the CSR. However, the proposed fusion scheme produces a closer value to the CSR. These average values indicates that the effectiveness of our technique is better as compared to the state-of-the-art techniques. So the fused images are attained by our proposed algorithm have important details compared to the considered SOTA techniques. The Table 2.6 and

Figure 2.8 show that the fused images obtained by our algorithm have lesser artifacts and noises as compared to the existing techniques.

Table 2.5: Quantitative comparisons of average values of the FMI_{dct}, N_{abf}, $SSIM_a$ and EPI_a on TNO database

Methods	CBF	WLS	CSR	RP	RP-SR	CS-MCA	CNN	Proposed
Avg.FMI_{dct}	0.26309	0.33102	0.34640	0.28210	0.27930	0.35841	0.35269	**0.36942**
Avg.N_{abf}	0.31727	0.21257	0.01958	0.22677	0.21444	0.06680	0.13280	**0.00563**
Avg.$SSIM_a$	0.59957	0.72360	0.75335	0.68424	0.67385	0.72964	0.71372	**0.76802**
Avg.EPI_a	0.57240	0.67837	**0.71130**	0.64488	0.63737	0.69154	0.68444	0.70909

Table 2.6: Quantitative comparisons of amount of noise added

Methods	CBF	WLS	CSR	RP	RP-SR	CS-MCA	CNN	Proposed
Image1	0.23167	0.14494	0.01494	0.18188	0.19922	0.05548	0.12243	**0.00709**
Image2	0.48700	0.16997	0.02199	0.32816	0.44373	0.08130	0.11717	**0.02188**
Image3	0.54477	0.21469	0.02070	0.26990	0.17392	0.06196	0.13342	**0.01031**
Image4	0.45288	0.22866	0.02378	0.31436	0.15666	0.07792	0.13154	**0.00665**
Image5	0.43257	0.19188	0.00991	0.23420	0.15975	0.03433	0.11558	**0.00022**
Image6	0.23932	0.22382	0.02296	0.18569	0.15037	0.06947	0.13074	**0.00107**
Image7	0.41779	0.15368	0.01514	0.15455	0.16687	0.07079	0.12248	**0.00670**
Image8	0.15233	0.23343	0.03404	0.35424	0.31220	0.09665	0.09363	**0.00468**
Image9	0.11741	0.17177	0.02371	0.11645	0.11665	0.07226	0.13313	**0.00088**
Image10	0.20090	0.22419	0.02013	0.29081	0.27369	0.07114	0.15679	**0.00193**
Image11	0.47632	0.20588	0.01022	0.09387	0.10909	0.04387	0.12432	**0.00143**
Image12	0.25544	0.22335	0.01545	0.21989	0.24685	0.05387	0.13585	**0.00822**
Image13	0.36066	0.19607	0.01888	0.27621	0.23131	0.06410	0.12659	**0.00747**
Image14	0.18971	0.20332	0.02036	0.16489	0.14370	0.06897	0.14908	**0.00265**
Image15	0.21509	0.20378	0.02207	0.18090	0.17750	0.07838	0.13583	**0.00081**
Image16	0.52783	0.30672	0.01936	0.33994	0.32261	0.06661	0.15255	**0.00978**
Image17	0.52887	0.31160	0.01561	0.21055	0.30008	0.06330	0.16254	**0.01268**
Image18	0.26649	0.25937	0.01499	0.20357	0.25116	0.05345	0.16153	**0.00808**
Image19	0.12582	0.16205	0.01379	0.16681	0.14794	0.04653	0.11416	**0.00245**
Image20	0.25892	0.18401	0.02574	0.32388	0.24738	0.08103	0.13493	**0.00160**
Image21	0.18091	0.25074	0.02745	0.15135	0.17251	0.09139	0.13453	**0.00165**

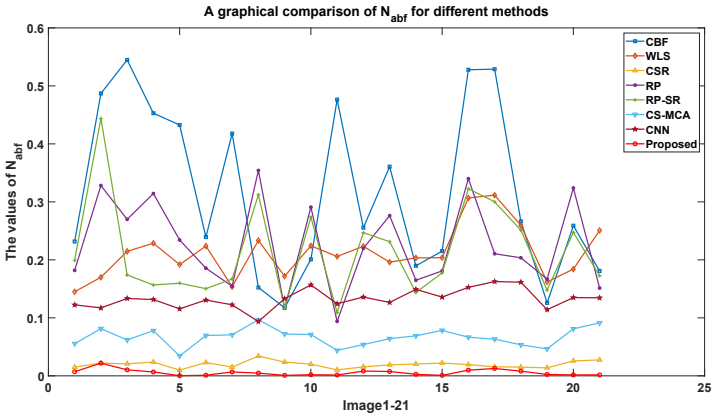


Figure 2.8: Quantitative comparisons of amount of noise added for different schemes.

2.3.5 Discussions and Future Works

Fusion of visible and infrared images is a challenging task in a vision-based system. This can be helpful to detect and track the moving objects in any surveillance system for the

target scene. Significant information extraction from the source images and propagating this information into the fused image without adding artifacts are quite difficult jobs in the image fusion process. In this chapter, we have proposed two pixel level image fusion schemes: fuzzy edge preserving intensity variation approach and weighted combination of maximum and minimum value selection strategy. The proposed algorithm results are validated qualitatively as well as quantitatively by comparing with its result those of the different state-of-the-art (SOTA) techniques. For fair evaluation, the SOTA techniques are implemented without altering the parameters. It may be observed that the proposed algorithms can retain maximum details by increasing the contrast and lesser artifacts in the fused image. Also, to know the efficiency of the proposed algorithms, we have performed a quantitative comparison among them. From Table 2.7, it may be observed that the proposed weighted combination of maximum and minimum value selection strategy for image fusion attain better accuracy against the proposed fuzzy edge preserving intensity variation approach in terms of all considered measures.

Table 2.7: Quantitative comparisons between the proposed fuzzy edge preserving intensity variation approach and weighted combination of maximum and minimum value selection strategy

Algorithms/ Quantitative measures	Fuzzy edge preserving intensity variation approach	Weighted combination of maximum and minimum value selection strategy
Avg.FMI_{dct}	0.31052	**0.36942**
Avg.N_{abf}	0.28250	**0.00563**
Avg.$SSIM_a$	0.60635	**0.76802**
Avg.EPI_a	0.66744	**0.70909**

The proposed fuzzy edge preserving intensity variation approach and weighted combination of maximum and minimum value selection strategy improve the visual contents of thermal sequences. However, all the parameters are fixed manually in the proposed fuzzy edge preserving intensity variation approach. An Expectation-Maximization algorithm can be used to fixing up those parameters. Again, in the proposed weighted combination of maximum and minimum value selection strategy, the weighted-average fusion approach improves the noise in the fused image. Using a statistical approach, one can reduce the noise in the fused image.

2.4 Conclusions

Two pixel level image fusion schemes have been proposed in this chapter. In the proposed fuzzy edge preserving intensity variation approach, we have investigated the spatial inter-dependency among the source images to generate the salient feature map with reduced artifacts. However, the salient feature map cannot preserve sufficient edge details. Therefore, the concept of the fuzzy edge is explored in the visible image to retain its edge details with significant contrast. The salient feature map with edge details generates a high contrast fused image with essential information. In the proposed weighted combination of maximum and minimum value selection strategy, a maximum selection strategy is explored to obtain the detailed feature map, which has the prominent details of the source images. However, the detail feature map cannot preserve the subtle details from the source images. Therefore, a minimum selection strategy is applied among the source images to generate an intermediate feature map with subtle details. A weighted combination of detail and intermediate feature map produces a fused image with standard features and reduced artifacts.

The proposed schemes are evaluated with challenging source pairs available at the benchmark *TNO* database. The efficacy of the proposed fuzzy edge preserving intensity variation approach is validated against eight state-of-the-art schemes. The efficiency of the proposed weighted combination of maximum and minimum value selection strategy is corroborated against the seven existing techniques. The performance of the proposed techniques is validated qualitatively and quantitatively in order to justify our findings. It is found that the proposed algorithms are attained higher accuracy against the considered state-of-the-art techniques.

Chapter 3

Integration of Multi-scale Features with Deep Learning Architecture for Feature Level Image Fusion

3.1 Introduction

A brief introduction of the existing feature level image fusion techniques are discussed in the Chapter 1. It can be concluded from the literature, that the sparse representation-based fusion schemes [33, 34, 35, 35] do not consider multi-scale decomposition strategy, which causes loss of many important details in the fused image. Further, in the deep learning-based fusion techniques, the extracted features are not fully utilized, and hence, may cause loss of information in the fused image. Such fused images are unable to be used for many higher order applications: object detection, object tracking, object recognition, etc. One of such examples are provided in Figure 3.1. Figures 3.1 (a) and (b) portrays a pair of visible and IR images considered for experimentation. The fused image obtained by an existing feature level fusion technique (deep neural network (DNN) [37]) is presented in Figure 3.1 (c). It may be observed from Figure 3.1 (c), that the fused image produced by the DNN technique is incapable of preserving significant details and produces a higher amount of artifacts. A zoomed portion of the fused image portray that, the region in the fused image does not carry meaningful information and is not suitable for human visible perception.

In this context, we have proposed two feature level image fusion schemes: integration

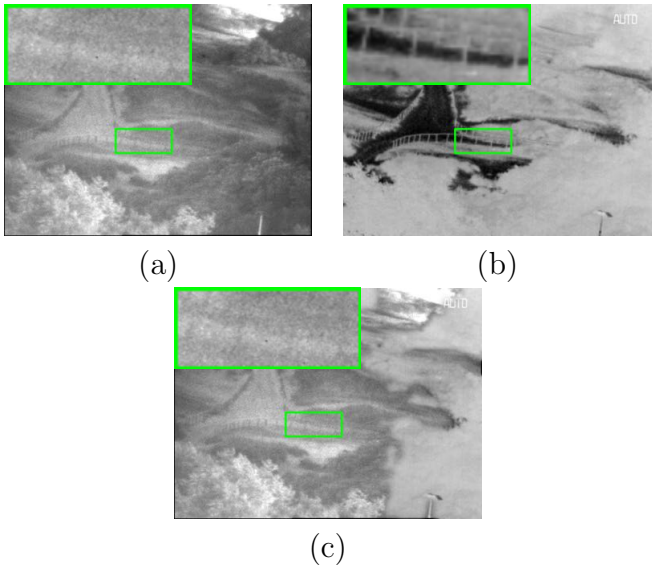


Figure 3.1: Visual analysis of (a) Visible image, (b) IR image, and (c) Fused image obtained by the DNN technique.

of bi-dimensional empirical mode decomposition with two streams VGG-16 and non-subsampled contourlet transform induced two streams ResNet-50 network in this chapter. In integration of bi-dimensional empirical mode decomposition with two streams VGG-16 technique, we have proposed the use of bi-dimensional empirical mode decomposition (BEMD) strategy integrated with an VGG-16 deep neural network architecture to retain the features of visible and infrared images in-depth at various levels in fused images. The fused image acquired by the proposed fusion strategy keeps the required details and strongly correlates with the source images. Further, in the proposed non-subsampled contourlet transform induced two streams ResNet-50 network technique, the integration of non-subsampled contourlet transform (NSCT) mechanism and ResNet-50 network is adhered to exploit the multi-scale, multi-directional, and shift-invariant details of the source images at low-frequency and high-frequency bands. The proposed novel fusion strategy generates the fused image that retains the background and object information from the source images efficiently.

The proposed schemes are evaluated on the benchmark *TNO* database. The efficacy of the proposed integration of bi-dimensional empirical mode decomposition with two streams VGG-16 scheme is corroborated against fifteen existing fusion schemes. Further, the performance of the proposed non-subsampled contourlet transform induced two streams ResNet-50 network algorithm is demonstrated against ten existing fusion schemes. We have used qualitative and quantitative analysis to confirm our findings. It is observed

that the proposed schemes provide better results than the existing SOTA techniques.

Further, the organization of this chapter is as follows. Section 3.2 describes the proposed feature level fusion schemes. The results and discussions with future works are presented in Section 3.3. Section 3.4 draws the conclusions of the proposed works.

3.2 Proposed Integration of Multi-scale Features with Deep Learning Architecture for Feature Level Image Fusion

In this chapter, we have proposed two feature level image fusion techniques: integration of bi-dimensional empirical mode decomposition with two streams VGG-16 and non-subsampled contourlet transform induced two streams ResNet-50 network. Here we assumed that, the I_s, s denotes the source images from various sensors, where $s \in \{1, 2\}$, 1 corresponds to the visible image, 2 corresponds to the IR image, and F is the fused image.

3.2.1 Proposed Integration of Bi-dimensional Empirical Mode Decomposition with Two Streams VGG-16

Usually IR and visible images have high uncertainty and may possess camera noise. Therefore, it is a quite challenging task to extract the meaningful features from both the images and propagate into a fused image with reduced artifacts. In this regard, we proposed a integration of bi-dimensional empirical mode decomposition with two streams VGG-16 technique that can extract multi-scale deep features from the source images. Here, the BEMD block is integrated with the VGG-16 network to preserve the deep multi-level visual details at various scales. It is observed that the proposed fusion strategy can retains significant visual details at a multi-level to produce a fused image with lesser artifacts. The graphical exposition of the proposed scheme is narrated in Figure 3.2. The schematic description of each block is described as follows.

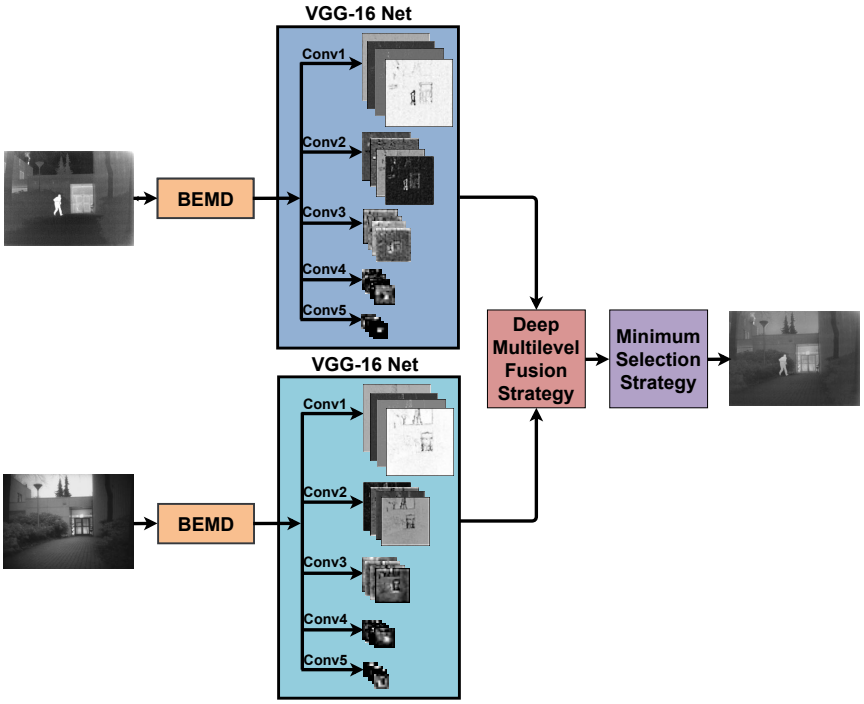


Figure 3.2: Block diagram of the proposed integration of bi-dimensional empirical mode decomposition with two streams VGG-16 scheme.

3.2.1.1 Bi-dimensional Empirical Mode Decomposition for Multi-Scale Feature Extraction

The empirical mode decomposition (EMD)[118] technique is popularly used in the signal and image processing domain to decomposes any signal into finite oscillatory components. It is an adaptive algorithm and relevant for stationary as well as non-stationary signal analysis. The extracted oscillatory components from the signal are named as intrinsic mode function (IMF). It is observed that the EMD mechanism plays an important role in one-dimensional signal analysis. The EMD mechanism is further extended and utilized for two-dimensional signal or image analysis and is known as bi-dimensional empirical mode decomposition (BEMD) [119]. The BEMD strategy extracts the $IMFs$ from the source images by utilizing the sifting process [120] and can be described as;

$$IMF_{s,n} = \{IMF_{s,1}, IMF_{s,2}, \cdots\cdots, IMF_{s,N-1}, R'_{s,N}\}; \tag{3.1}$$

$$\forall n = 1, 2, \cdots\cdots, N-1, N.$$

where $IMF_{s,n}$ indicates the $IMFs$ for the images of visual and thermal sensors. n indicates the number of $IMFs$ and $R'_{s,N}$ denotes the residue bands of source images.

3.2.1.2 VGG-16 Net for Fusion of Intrinsic Mode Functions

In the proposed scheme, we have introduced the BEMD strategy to decompose the source images into N numbers of intrinsic mode functions at multi-scale with different frequency bands. To retain the maximum details in the fused image with reduced noise, it is necessary to extract the deep features at different levels from the $IMFs$ and combine them accurately. Therefore, we have proposed a unique deep multi-level fusion strategy with VGG-16 (deep learning architecture) [121] that provides the multi-scale and multi-level visual characteristics of the considered scene. The considered VGG-16 architecture consists of convolutional layers, max-pooling layers, and a rectified linear unit (ReLU) as activation function with five convolutional blocks. The convolutional layers are used to retain the spatial information of the source images, and max-pooling layers are used for the down-sampling operation. The presence of the ReLU function in the network makes it faster and efficient.

In the proposed algorithm, we have used $IMF_{sn}(x, y)$ except residual bands of source images to extract the deep features as shown in Figure 3.2. Let us assume that Θ_s^i deep features are extracted by the VGG-16 network with i^{th} convolutional block, $i \in \{1, 2, 3, 4, 5\}$. $\Theta_s^i(x, y)$ indicates the contents at position (x, y). The visual and the IR images $IMFs$ are given as the input to the two streams of the VGG-16 net to retain the deep features. Subsequently, at each block, we have utilized the sum of the absolute difference (SAD) operator to map from features space to image space named as activity level map A_s^i and can be given as;

$$A_s^i(x, y) = SAD(\Theta_s^i(x, y)). \tag{3.2}$$

To make the proposed scheme prosperous to mis-registration, we have considered a center sliding window $w \times w$ in the A_s^i. A block-based average (BBA) operator in the center sliding window is used to obtain the action level map $\tilde{A}_s^i$ and can be calculated as;

$$\tilde{A}_s^i(x, y) = \frac{\sum_{p=-\frac{w-1}{2}}^{\frac{w-1}{2}} \sum_{q=-\frac{w-1}{2}}^{\frac{w-1}{2}} A_s^i(x+p, y+q)}{w^2}. \tag{3.3}$$

For the larger value of w, the proposed algorithm is more prosperous to mis-registration. However, at the same time, small-scale details may be lost, which are essential for multi-

modal image fusion. Hence, in proposed work, we have kept the size of the center sliding window as 3×3.

To precisely preserve the complementary information from the source images, we have developed an action weight map W_s^i by using the normalization operator in the $\tilde{A}_s^i$. The W_s^i in the range [0,1] can be determined as;

$$W_s^i(x,y) = \frac{\tilde{A}_s^i(x,y)}{\sum_{m=1}^{2} \tilde{A}_m^i(x,y)}. \tag{3.4}$$

As we know, the max-pooling layer with stride of 2 in the VGG-16 architecture reduces the size of the input feature to 1/2 times. Hence, the bi-cubic interpolation is utilized in the W_s^i to generate the intermediate weight map $\tilde{W}_s^i$ where the size of $\tilde{W}_s^i$ same as the source images size.

The intermediate feature maps IFM^i are generated from the source images and intermediate weight maps to retain the high strength details and remove the low strength details. Considering five layer of two streams VGG-16 network, we have five pairs of intermediate weight maps, and for each pair of intermediate weight maps, the intermediate feature map can be calculated as;

$$IFM^i(x,y) = \sum_{z=1}^{s} \tilde{W}_z^i(x,y) \times I_z(x,y). \tag{3.5}$$

The detail feature map DFM is obtained by using maximum selection strategy among these intermediate feature maps to preserve sharp details and can be calculated as;

$$DFM(x,y) = max[IFM^i(x,y)], \tag{3.6}$$

likewise, we acquired various detail feature maps from the corresponding $IMFs$.

3.2.1.3 Fused Image Generation

To generate the fused image F, we have utilized a minimum selection strategy among these detail feature maps to preserve the standard information and reduce the redundant data. The fused image is obtained as;

$$F(x,y) = min[DFM^n(x,y)], \tag{3.7}$$

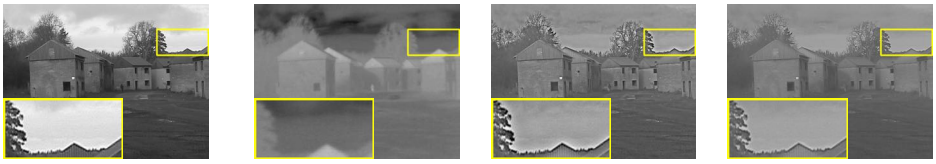
where n indicates the number of $IMFs$ pairs.

3.2.2 Proposed Non-subsampled Contourlet Transform Induced Two Streams ResNet-50 Network

It is observed from the state-of-the-art techniques, the non-subsampled contourlet transform (NSCT) decomposition mechanism has considered a transform-domain-based feature extraction strategy to extract multi-scale and shift-invariant features from infrared and visual images. Further either low frequency [42] or high frequency [40] coefficients are used as the input to the deep CNN architecture for extraction of deep multi-scale features which are used for fusion. It is to be noted that any fusion architecture will be incomplete without the utilization of both low as well as high-frequency features. Hence such approaches are found to be not providing a significant accuracy in the fusion process.

In addition to this; one of the recent fusion architecture reported in [42] uses the deep features extracted from low-pass coefficients of NSCT which are not rich in edge details and hence unable to preserve the structural information in the fused image. Further, the band-pass coefficients are found to be fused using a modulus maximum selection strategy to propagate artifacts into the fused images. The performance of the said techniques in the *TNO* database as reported with measures: amount of noise added to the fused images by the fusion process to be 0.14742, average structural similarity index measure ($SSIM_a$) to be 0.74004 and the average edge-preserving index (EPI_a) to be 0.71303. It shows a poor performance of the said [42] technique. This is mainly due to the loss of high-frequency details in the fused images and the use of modulus maximum selection strategy that propagated artifacts in the fused image. Similarly, the DL technique [40] which uses the high-frequency coefficients in the deep learning architecture are found to propagate a higher amount of noise in the fused image. It may be observed that the said technique reported the amount of noise added to the fused images by the fusion process measure to be 0.00267 for the *TNO* database. The poor performance of the said technique is mainly due to the fact that noise propagates with high-frequency details. Also, it decreases the accuracy of the fusion process. The results reported by the said approaches are shown in Figure 3.3. It can be clearly visible that use of low frequency coefficients by [42] as shown in Figure 3.3 (c) resulted in contour effects along the edges. Similarly, the use of high frequency coefficients [40] for fusion has created a large amount of noise across the edges

shown in Figure 3.3 (d). Hence this reduces the accuracy of the said approaches.



(a) Visible image (b) IR image (c) [42] (d) [40]

Figure 3.3: Visual analysis of (a) Visible image, (b) IR image, (c) Fused image obtained by the ResNet-152 based technique and (d) Fused image obtained by the DL technique

In the proposed work, we tried to address the above mentioned challenges by designing a fusion technique that takes care of both low as well as high-frequency coefficients of the transform domain features. It also may be concluded from the above analysis that low-frequency coefficients, in conjunction with a high-frequency coefficients, can produce considerable improvement in the fusion accuracy.

In the proposed work, we tried different multi-scale decomposition mechanisms with the two-stream ResNet-50 network and found that the Non-subsampled contourlet transform (NSCT) induced two-stream ResNet-50 architecture to be efficient for multi-modal image fusion. The steps of the proposed fusion mechanism are described as follows. The motivation and contribution behind considering the NSCT mechanism with two-stream ResNet-50 residual network and proposed novel fusion strategy is as follows:

(1) The proposed algorithm relies on a non-subsampled contourlet transform (NSCT) decomposition mechanism to avoid the frequency aliasing problem and enhance the directional selectivity as well as shift-invariance of the details from the thermal and visual images.

(2) The proposed technique utilizes both high frequency and low-frequency coefficients in two parallel Res-Net-50 networks with residual connections to extract the multi-scale deep features which essentially characterize the subtle and detailed information of images captured from visual and infrared sensors.

(3) It may be observed that the deeper blocks of the ResNet-50 network gradually learn more complex features and hence provides an improved performance.

(4) The lower blocks of the ResNet-50 network can learn and obtain the low-level local features such as edges, colors, and textures, while the deeper blocks learn and obtain high-level global features like objects and events.

(5) The NSCT mechanism followed by the two-stream ResNet-50 network can preserve the deep features at multi-scale and multi-direction with different levels that are essential for IR and visual image fusion.

(6) The proposed fusion strategy precisely retains the significant visual information from the source images that generate a fused image with maximum details and reduced artifacts.

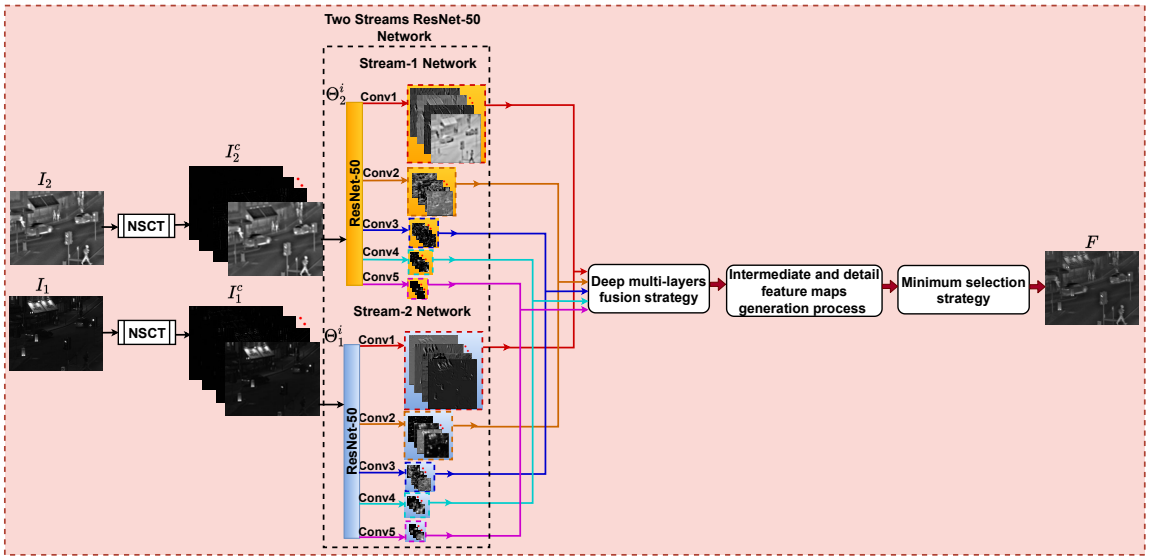


Figure 3.4: Block diagram of the proposed non-subsampled contourlet transform induced two streams ResNet-50 network scheme.

3.2.2.1 Source Images Decomposition

In this stage, the images from various sensors are decomposed into low and high-frequency coefficients using the NSCT [122, 29, 24]. In contourlet transform, the directional filter banks (DFBs) and LP filters are used for directional and multi-scale decomposition. To obtain shift-invariance and exclude the frequency alias of the contourlet transform, Cunha *et al.* [123] developed the non-subsampled contourlet transform (NSCT). This NSCT comprises non-subsampled pyramid filter banks (NSPFBs) and non-subsampled directional filter banks (NSDFBs).

To obtain the multi-scale decomposition of the images of different sensors, the NSPFB is utilized. At each scale, the NSDFB is employed to divide and pass high-frequency coefficients in several directions. One low-frequency and one high-frequency coefficient can be obtained for the source images at each NSPFB decomposition level. The ensuing

NSPFB levels are performed to break down the low-frequency coefficient iteratively that preserves the vital details in the image. Therefore, NSPFB can provide $j+1$ coefficients consisting of one low-frequency coefficient and j high-frequency coefficients with the same size as the source image. Here, j indicates the decomposition levels. Then, the NSDFB is utilized to decompose the high-frequency coefficients at each scale, providing directional components of the same size as the source image. The NSDFB is designed by using the directional fan filter banks [29] which are two-channel non-subsampled filter banks. In this work, NSPFB and NSDFB are performed using the 'maxflat' [124] and 'dmaxflat7' [124] filters. We have used the 'maxflat' filter for pyramidal decompositions, and the 'dmaxflat7' filter of order 7 is for directional coefficients. This pair of filters provides better accuracy and avoids smearing of image details. We have kept the NSCT decomposition level of 4 to preserve maximum detail from the sources. It is found that the NSCT with decomposition level 4 retains small-scale and large-scale features from the sources which are suitable for image fusion.

From the above discussion, as we can see, the NSCT has the characteristics of contourlet transform and has shift-invariance properties. In the NSCT, the size of the different coefficients are the same; hence the relation among the coefficients are easily found which is useful in designing the rules for image fusion. Also, the effect of mis-registration reduced efficiently in the fused image [29]. Therefore, the NSCT is more appropriate for image fusion.

3.2.2.2 Two Stream Network for Feature Extraction and Fusion of the Coefficients

Conventional neural networks have limited layers, which cause the loss of information during feature extraction. In order to retain maximum information in the fused images, deeper neural networks may be employed. However, due to the vanishing gradient [125] and degradation [126] problem, it is challenging to train the deep neural network. In [127], the authors put forth a residual network to cater to this problem. With the residual representation and the shortcut connections, this network is easier to optimize and provides better accuracy by increasing the depth. Further, the low and high-frequency coefficients must be separated and unequally needs to be weighted through the residual networks to properly characterize the deep features corresponding to the infrared and visual images.

For the low and high-frequency coefficients, we have used a two-stream ResNet-50 network, which extracts the deep multi-layer features as shown in Figure 3.4. The proposed two-stream network runs parallel, where one stream runs on the low-frequency coefficients of the NSCT and another stream runs on the high-frequency coefficients of the NSCT. In both streams, ResNet-50 is used. This network consists of 50 weight layers which include five convolutional blocks (conv1, conv2, conv3, conv4, conv5) and is trained by ImageNet [128]. The lower block of the ResNet-50 network learns features like edges or colors of the image, and the higher block of the network learns features like objects and events. In this work, the corresponding coefficient pairs are given to the deep neural network, separately. This network allows coefficient information to be passed directly to the subsequent layers, removing the same information and highlighting minor changes. Therefore, this network explores the diverse details from coefficients that enhance the efficacy of the proposed scheme.

The residual block of the deep neural network consists of convolutional layers, batch normalization layers, and a rectified linear unit (ReLU) as an activation function. Convolutional layers are employed to obtain the spatial information of the source images using convolutional kernels. For faster learning rates and properly initializing the neural network, the batch normalization operation is performed after each convolution layer. The use of the ReLU function introduces non-linearity that makes the network faster and more efficient. After constructing the residual block, the stacked residual blocks depict the ResNet-50 network [127]. The ResNet-50 network has wide applications in the field of computer vision, however, it has not been used much in image fusion techniques.

We can represent the residual block of the deep neural network as;

$$\mathcal{Q}_t = RBO(P_t, \mathcal{W}_t) + P_t. \tag{3.8}$$

$$P_{t+1} = \Re(\mathcal{Q}_t). \tag{3.9}$$

$$\Re(\mathcal{Q}_t) = \begin{cases} \mathcal{Q}_t & \mathcal{Q}_t > 0 \\ 0 & \mathcal{Q}_t \leq 0 \end{cases}. \tag{3.10}$$

where P_t and $\mathcal{Q}_t$ indicates the input and output feature of the t^{th} residual block. $\mathcal{W}_t = \{W_{t,r}|_{1\leq r\leq \mathbb{R}}\}$ is a set of weights and biases, and $\mathbb{R}$ is the number of layers associated with the t^{th} residual block. $RBO(\cdot)$ represent the residual block operation, *e.g.*, a stack of convolution, batch normalization and ReLU layers and $\Re(\cdot)$ is a ReLU function.

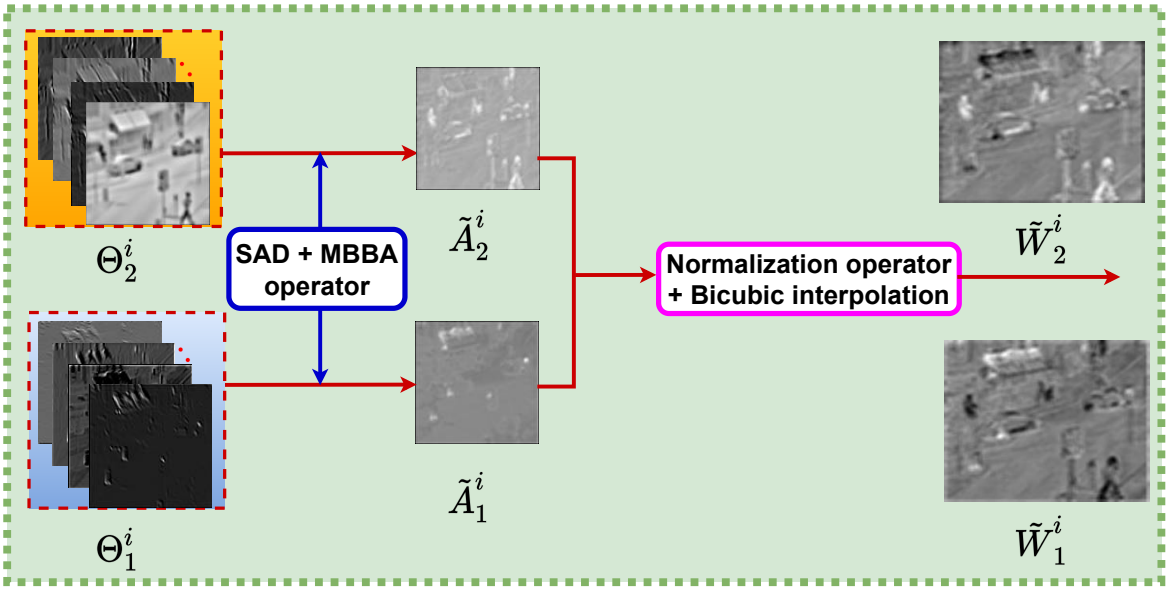


Figure 3.5: Block diagram of the proposed deep-multi layers fusion strategy.

We have proposed a novel fusion strategy for the multi-layer deep features from all the convolutional blocks of the ResNet-50 network that preserve as much information as possible. The proposed deep multi-layers fusion strategy is shown in Figure 3.5. Let us assume the I_s^c represent the coefficients pairs and the Θ_s^i denotes the deep features extracted by the ResNet-50 network with i^{th} convolutional block, $i \in \{1, 2, 3, 4, 5\}$. $\Theta_s^i(x, y)$ represent the information at the position (x, y) in the deep features. Initially, a similar kind of coefficients pair are given to the ResNet-50 network individually to obtain the Θ_s^i. Then, at each block, to retain the edge details and geometric structure from the deep features, we have proposed an activity level map construction process where the sum of the absolute difference SAD operator is used pixel-by-pixel basis among these feature maps. Here, we determined the absolute difference between the succeeding feature maps in the multi-layer deep features and combined them to generate the activity level map A_s^i and can be calculated as;

$$A_s^i(x, y) = SAD(\Theta_s^i(x, y)). \tag{3.11}$$

To make the proposed scheme insensitive to mis-registration, the action level map $\tilde{A}_s^i$ is obtained by considering a center sliding window $w \times w$ in the A_s^i. Further, A_s^i is zero-padded with $\frac{w-1}{2}$ numbers of rows and columns. Also, MBBA operator is considered in this window to obtain the $\tilde{A}_s^i$ and can be calculated as;

$$\tilde{A}_s^i(x,y) = \frac{\sum_{p=-\frac{w-1}{2}}^{\frac{w-1}{2}} \sum_{q=-\frac{w-1}{2}}^{\frac{w-1}{2}} A_s^i(x+p, y+q)}{w^2}. \tag{3.12}$$

With a larger value of w, the fusion method is excepted to be more robust to mis-registration, but some small-scale details may be lost simultaneously. As the small-scale details are frequently required in the multi-modal image fusion, it is more suitable to choose a smaller value of the w. For the fusion process, we set the dimensions of the center sliding window as 3×3. It is found that the center sliding window with a size of 3×3 explores the spatial dependency among neighborhood pixels very well to preserve small-scale details from the sources against the center sliding window with size of 5×5, 7×7 and 9×9.

3.2.2.3 Weight Maps Generation

The IR and visible images provide complementary information, i.e., the information present in one image might not be present in the other. Generally, the IR images provide thermal radiation information where the object is identified, but the background information is insufficient. However, the visible image highlights the background information but fails to provide the object's information. We need to integrate objects and background information from the two source images into one fused image. To this end, we assign low weight to pixels with less significant information and vice-versa. Hence, in this work, we have proposed a weight map construction process where the normalization operator is used in the $\tilde{A}_s^i(x,y)$ to obtain the action weight map $W_s^i(x,y)$. The $W_s^i(x,y)$ is considered to be in the range of [0,1] and can be represented as;

$$W_s^i(x,y) = \frac{\tilde{A}_s^i(x,y)}{\sum_{m=1}^{2} \tilde{A}_m^i(x,y)}. \tag{3.13}$$

In the conv1 block of ResNet-50 network, 7×7 convolutional layer with stride of 2

is used which reduces the size of the input image to 1/2 times of the actual dimensions of the input image. The max-pooling layer is used in the ResNet-50 network after the conv1 block with stride of 2 which is a kind of sub-sampling operation. This layer reduces the size of deep features to 1/2 times of the actual dimensions of the input deep features. Again, at the beginning of the rest of the blocks such as conv3, conv4, and conv5, 1×1 convolutional layer with stride of 2 is used which further reduces the size of the deep features to 1/2 times of actual dimensions. Hence, the intermediate weight map $\tilde{W}_s^i$ is obtained by utilizing the bicubic interpolation which resizes the W_s^i into the source image size.

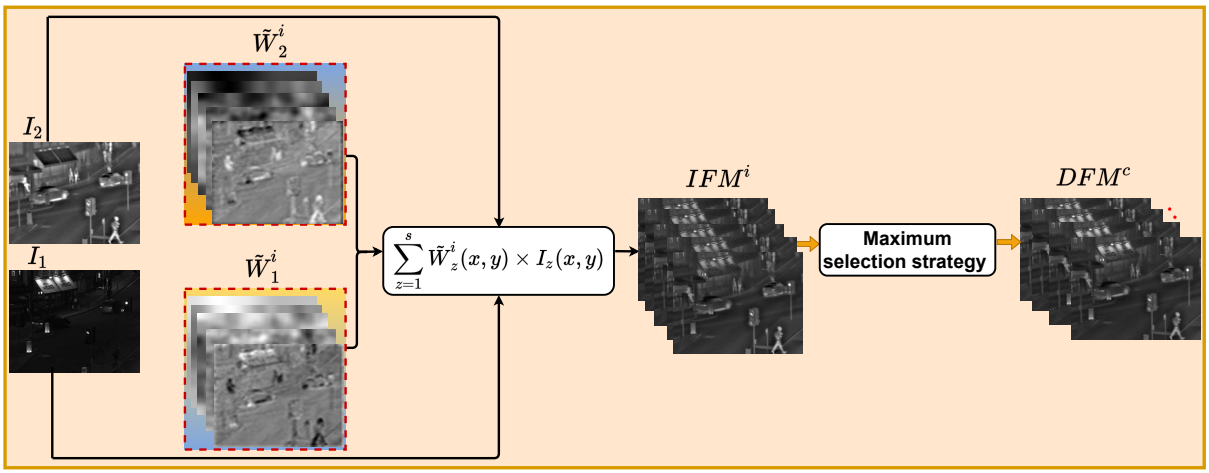


Figure 3.6: Block diagram of the proposed intermediate and detail feature maps generation process.

The intermediate feature map IFM^i can be obtained by using the source images and the intermediate weight maps are shown in Figure 3.6 to preserve the high strength salient features and to discard the low strength salient features of the source images. Now we have five pairs of intermediate weight maps $\tilde{W}_s^i$. For each pair of $\tilde{W}_s^i$, the intermediate feature map is calculated as;

$$IFM^i(x,y) = \sum_{z=1}^{s} \tilde{W}_z^i(x,y) \times I_z(x,y). \tag{3.14}$$

Finally, the detail feature map DFM is determined by applying the maximum selection strategy among these five intermediate feature maps are presented in Figure 3.6 to preserve the sharp features and can be given as;

$$DFM(x,y) = max[IFM^i(x,y)], \tag{3.15}$$

similarly, we obtained several detail feature maps from all the coefficient pairs.

3.2.2.4 Fused Image Generation

The detail feature maps we obtained by the above process have the sources' textural details and thermal radiation information. The thermal radiation information is generally characterized by pixel intensities which make the objects easily identifiable. Generally, the textural details are mainly characterized by the gradients and provide detailed information for the scene. For the complete description of the scene, objects, and detailed information of the target scene is highly essential for the fused image F. Therefore, to get the objects as well as detailed information in the F, the minimum selection strategy is applied among these detail feature maps from the various coefficient pairs, which conserve the common information and decrease the redundant information and can be given as;

$$F(x,y) = min[DFM^{c}(x,y)], \tag{3.16}$$

where c denotes the number of coefficient pairs.

3.3 Results and Discussions

The proposed algorithms are implemented on a 16 GB RAM equipped with *Core* i7 system, 1.5 MB L2 cache. The proposed schemes are tested on all the source image pairs available on the *TNO* benchmark database [2].

In this section, the performance of the proposed algorithms are validated qualitatively as well as quantitatively. To evaluate the performance of the proposed algorithms, we have used four quantitative measures: mutual information for the discrete cosine features (FMI_{dct}) [50], amount of artifacts added during the fusion process (N_{abf}) [51], average structure similarity index ($SSIM_a$) [52], and average edge preservation index (EPI_a) [53]. The achievement of the proposed integration of bi-dimensional empirical mode decomposition with two streams VGG-16 technique is verified by comparing the results obtained by it with those of the recently developed fifteen existing state-of-the-art fusion techniques: cross bilateral filter (CBF) [116], weighted least square (WLS) [117], convolutional sparse representation (CSR) [34], ratio of low-pass pyramid (RP) [24], RP with sparse representation (RP-SR) [24], latent low-rank representation (LatLRR) [129], morphological

component analysis based on convolutional sparsity (CS-MCA) [35], Fuzzy edge, Joint SR with saliency detection (JSRSD) [130], and saliency detection (SLD)[31], convolutional neural network (CNN) [36], deep neural network (DNN) [37], Fusion based on generative adversarial network (FusionGAN) [131], image fusion based on CNN (IFCNN) [132], and residual fusion network (RFN) [41]. To justify the efficiency of the proposed non-subsampled contourlet transform induced two streams ResNet-50 network technique, we compared it against with ten state-of-the-art image fusion techniques: cross bilateral filter (CBF) [116], weighted least square (WLS) [117], convolutional sparse representation (CSR) [34], ratio of low-pass pyramid (RP) [24], RP with sparse representation (RP-SR) [24], latent low-rank representation (LatLRR) [129], morphological component analysis based on convolutional sparsity (CS-MCA) [35], convolutional neural network (CNN) [36], deep neural network (DNN) [37], and deep learning based fusion (DL) [40].

3.3.1 Qualitative illustration of Integration of Bi-dimensional Empirical Mode Decomposition with Two Streams VGG-16

The original images acquired from the visual and the thermal sensors along with the results obtained by the proposed and different considered state-of-the-art techniques: CBF, RP, RP-SR, Fuzzy edge, RFN, and DNN are presented in Figure 3.7. It may be observed that the resultant images procured by the different techniques used for comparison: CBF, RP, RP-SR, and Fuzzy edge have produced many artifacts and cannot retain significant details in the fused image, as shown in the red rectangle highlighted region on different images. The outcomes of the RFN technique have blurred details with more noise. Due to ringing artifacts around the edge details, the significant features are not clearly visible or highlighting non-required details in the fused image by the DNN technique. However, the results obtained by the proposed technique have maximum details with lesser artifacts.

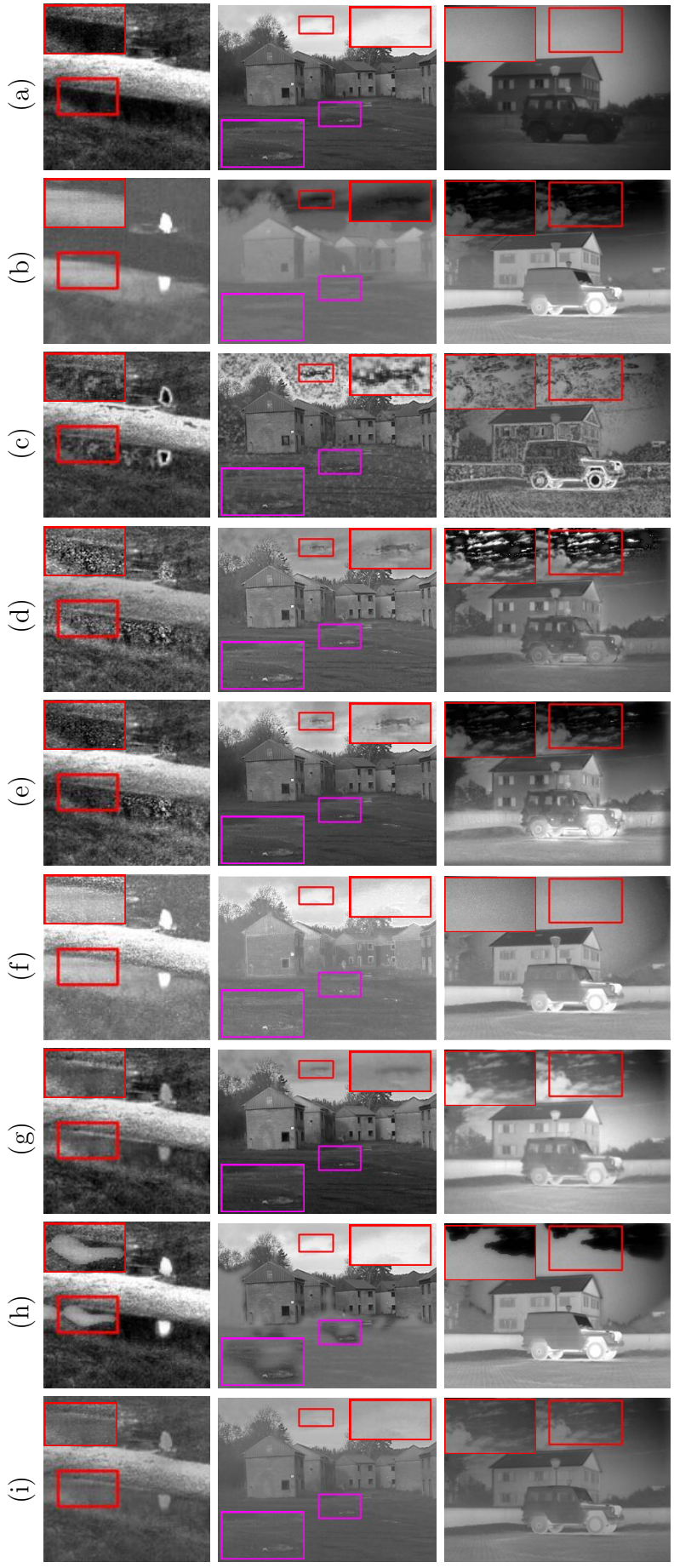


Figure 3.7: Visual analysis of results on Bench, Octec, and Marne images (from left to right). From top to bottom: (a) Visible images, (b) IR images, fused images obtained by (c) CBF, (d) RP, (e) RP-SR, (f) Fuzzy edge, (g) RFN, (h) DNN and (i) proposed integration of bi-dimensional empirical mode decomposition with two streams VGG-16 scheme.

3.3.2 Quantitative comparison of Integration of Bi-dimensional Empirical Mode Decomposition with Two Streams VGG-16

The evaluation of fusion performance is difficult due to irrelevant variation in visual demonstrated fusion results obtained by different fusion schemes. Therefore, in this paper we have used the four most suggested fusion metrics: mutual information for the discrete cosine features (FMI_{dct}) [50], amount of artifacts added during the fusion process (N_{abf}) [51], average structure similarity index ($SSIM_a$) [52], and average edge preservation index (EPI_a) [53].

Table 3.1: Quantitative comparisons of average values of the FMI_{dct}, N_{abf}, $SSIM_a$ and EPI_a on TNO database

Quantitative measurements /Algorithms	Avg.FMI_{dct}	Avg.N_{abf}	Avg.$SSIM_a$	Avg.EPI_a
CBF [116]	0.26309	0.31727	0.59957	0.57240
WLS [117]	0.33102	0.21257	0.72360	0.67837
CSR [34]	0.34640	0.01958	0.75335	0.71130
RP [24]	0.28210	0.22677	0.68424	0.64488
RP-SR [24]	0.27930	0.21444	0.67385	0.63737
LatLRR [129]	0.33817	0.01596	0.76486	0.76223
CS-MCA [35]	0.35841	0.06680	0.72964	0.69154
Fuzzy edge	0.31052	0.28250	0.60635	0.66744
JSRSD [130]	0.14253	0.34657	0.54127	0.47473
SLD [31]	0.27030	0.13430	0.72897	0.66774
CNN [36]	0.35269	0.13280	0.71372	0.68444
DNN [37]	0.36658	0.02324	0.70852	0.68552
FusionGAN [131]	0.36335	0.06706	0.65384	0.68470
IFCNN [132]	0.37378	0.17959	0.73186	0.73767
RFN [41]	0.29669	0.07288	0.69949	0.68864
Proposed	**0.39962**	**0.00149**	**0.77671**	**0.77909**

Table 3.1 encapsulates the average quantitative measures of the proposed and the state-of-the-art fusion schemes where the best values are indicated in bold. From this Table, it may be observed that the proposed scheme attained higher accuracy in terms of all considered measures against fifteen existing fusion algorithms. Also, the fused images obtained by the proposed scheme include important visual content with maximum edge details due to the higher average value of FMI_{dct} and EPI_a against the existing fusion schemes. Further, the fused images acquired by the proposed algorithm strongly correlate with source images and contain fewer artifacts as compared to the state-of-the-art techniques because of the best average value of $SSIM_a$ and N_{abf}.

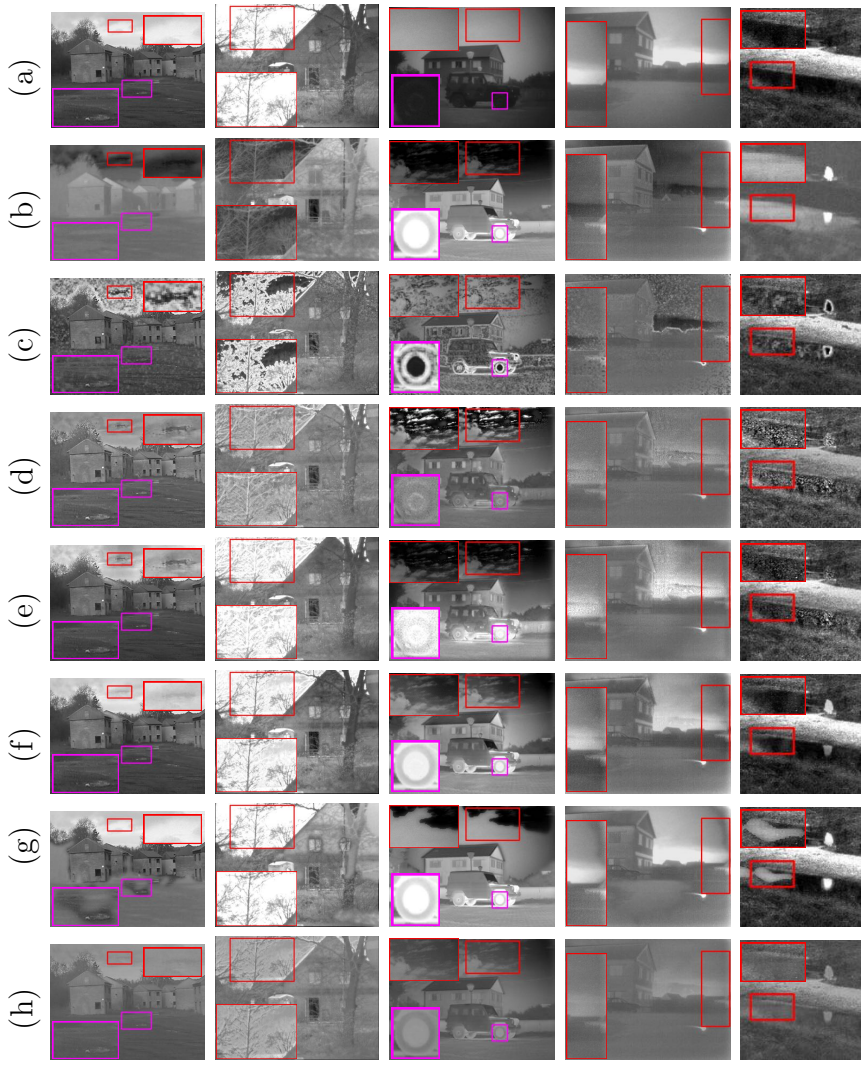


Figure 3.8: Visual analysis of results on Octec, Man in front of house, Marne, Movie 18, and Bench images (from left to right). From top to bottom: (a) Visible images, (b) IR images, fused images obtained by (c) CBF, (d) RP, (e) RP-SR, (f) CNN, (g) DNN, and (h) proposed non-subsampled contourlet transform induced two streams ResNet-50 network scheme.

3.3.3 Qualitative illustration of Non-subsampled Contourlet Transform Induced Two Streams ResNet-50 Network

The paired source and fused images acquired from the proposed method and other state-of-the-art techniques are shown in Figure 3.8. The performance of fusion of the proposed scheme is evaluated on few IR and visible image pairs taken from the *TNO* benchmark database: Octec, Man in front of house, Marne, Movie 18 and Bench.

Figure 3.8 (a) and Figure 3.8 (b) denotes the visible and the IR images, respectively. It may be observed that from Figure 3.8 (c), (d) and (e) the fused images generated by CBF, RP, and RP-SR contain more noise and the details are not clearly visible. The

resulting images obtained by CNN are shown in Figure 3.8 (f). It can be seen that these techniques have produced more artifacts. Also, the image details are blurred. The detailed information in images after fusion are shown in Figure 3.8 (g). These are obtained by DNN and the details are not clear because of the ringing artifacts around the features. On the contrary, the fused images are shown in Figure 3.8 (h) obtained by the proposed scheme having fewer artifacts, preserves more detailed information, and looks more natural as compared to these existing techniques.

3.3.4 Quantitative comparison of Non-subsampled Contourlet Transform Induced Two Streams ResNet-50 Network

Evaluation of the performance of fusion techniques is difficult as the ground truth is not always available for most challenging scenes. It is observed that in most of the literature cited the standard quantitative evaluation measures: mutual information for the discrete cosine features (FMI_{dct}) [50], amount of noise added to fused images by the fusion process (N_{abf}) [51], average structure similarity index ($SSIM_a$) [52] and average edge preservation index EPI_a [53]. Hence considering the importance of this work, we also adhered to the same.

The values of FMI_{dct}, N_{abf}, $SSIM_a$ and EPI_a are reported in Table 3.2 - 3.5 and the same is graphically represented in Figure 3.9 - 3.12 for the proposed and the ten state-of-the-art techniques: cross bilateral filter (CBF) [116], weighted least square (WLS) [117], convolutional sparse representation (CSR) [34], ratio of low-pass pyramid (RP) [24], RP with sparse representation (RP-SR) [24], latent low-rank representation (LatLRR) [129], morphological component analysis based on convolutional sparsity (CS-MCA) [35], convolutional neural network (CNN) [36], deep neural network (DNN) [37], and deep learning based fusion (DL) [40] for challenging image pairs, where values in bold indicate the best ones. It can be seen from Table 3.2 and Figure 3.9, the proposed method has the best value of FMI_{dct} as compared to the state-of-the-art techniques for the camp image to movie 18 image. However, the proposed algorithm produces the FMI_{dct} value for the sandpath, soldier behind smoke, and soldier in trench images are acceptable. From Table 3.3 and Figure 3.10, the N_{abf} values of fused images are obtained by the proposed technique are satisfactory for all the images excluding marne and soldier behind smoke images. The $SSIM_a$ values of our proposed technique shown in Table 3.4 and Figure

3.11, is best as compared to the state-of-the-art techniques excluding the DL technique for the street, bunker, soldier behind smoke, and soldier in trench images. However, the proposed technique produces comparable results to the DL. As compared with all the state-of-the-art techniques, the proposed algorithm obtains the best values in EPI_a are shown in Table 3.5 and Figure 3.12, for all the fused images.

From Table 3.6, it is clear that the average values of FMI_{dct}, N_{abf}, $SSIM_a$ and EPI_a metrics, that are obtained by the proposed algorithm for all the IR and visible image pairs, have the best average values and are indicated in bold against the state-of-the-art techniques. Observing these values, it is clear that the fused images generated by the proposed approach have better performance of image fusion than the state-of-the-art techniques. Also, the proposed method preserves sufficient structural information and features because of the best average values in $SSIM_a$ and FMI_{dct}. Furthermore, for the same reason, the fused images obtained by the proposed method are relatively more natural and contain fewer artifacts because of the best average values of EPI_a and N_{abf}.

Table 3.2: Quantitative comparisons of mutual information for the discrete cosine features

Methods / Images	CBF	WLS	CSR	RP	RP-SR	LatLRR	CS-MCA	CNN	DNN	DL	Proposed
Camp (CM)	0.24495	0.26995	0.27958	0.23513	0.23373	0.33541	0.28777	0.26425	0.28822	0.37289	**0.37352**
Street (ST)	0.28458	0.33872	0.37597	0.26278	0.21694	0.36015	0.36819	0.35397	0.39065	0.39408	**0.40121**
Man in front of house (MFT)	0.25719	0.33967	0.37682	0.30276	0.29940	0.35512	0.39458	0.39683	0.36878	0.43083	**0.43164**
Airplane in trees (AIT)	0.13701	0.21218	0.20117	0.18999	0.18390	0.20096	0.22734	0.23828	0.26310	0.28504	**0.28760**
Bunker (BR)	0.39827	0.37730	0.42364	0.35298	0.35935	0.36187	0.43206	0.41950	0.44599	0.44724	**0.44902**
Kaptein 1123 (KP1123)	0.22292	0.29200	0.27340	0.21496	0.19783	0.32829	0.27580	0.27313	0.28839	0.38330	**0.38710**
Kaptein 1654 (KP1654)	0.22987	0.29371	0.28996	0.24721	0.23381	0.32103	0.30758	0.28810	0.30120	0.38961	**0.39118**
Lake (LE)	0.30142	0.35993	0.38601	0.30381	0.29754	0.35490	0.40938	0.40580	0.40987	0.42693	**0.42949**
Men in doorway (MID)	0.28895	0.33821	0.38661	0.30284	0.30309	0.35787	0.40336	0.39310	0.40295	0.43667	**0.43714**
Marne (ME)	0.18351	0.33048	0.28026	0.18682	0.20946	0.33252	0.27684	0.30218	0.27160	0.39157	**0.39273**
Movie 1 (MV1)	0.17596	0.31727	0.29606	0.22095	0.20910	0.30702	0.30989	0.32215	0.33448	0.36826	**0.37218**
Movie 18 (MV18)	0.22016	0.30172	0.27818	0.24503	0.23262	0.32068	0.26703	0.27907	0.24446	0.38474	**0.38569**
Sandpath (SP)	0.26415	0.26049	0.29298	0.20958	0.20050	0.31613	0.33588	0.25211	0.29544	**0.37606**	0.37599
Soldier behind smoke (SBS)	0.20102	0.37284	0.42015	0.26611	0.26560	0.35426	0.41893	0.41351	**0.54888**	0.44377	0.44496
Soldier in trench (SIT)	0.31521	0.41898	0.42609	0.39744	0.41310	0.37319	0.42641	0.44681	0.44870	**0.45949**	0.45838

Table 3.3: Quantitative comparisons of amount of noise added

Methods / Images	CBF	WLS	CSR	RP	RP-SR	LatLRR	CS-MCA	CNN	DNN	DL	Proposed
CM	0.23167	0.14494	0.01494	0.18188	0.19922	0.01720	0.05548	0.12243	0.02976	0.00013	**0.00012**
ST	0.48700	0.16997	0.02199	0.32816	0.44373	0.02922	0.08130	0.11717	0.05862	0.00381	**0.00271**
MFT	0.23932	0.22382	0.02296	0.18569	0.15037	0.00873	0.06947	0.13074	0.02096	0.00099	**0.00039**
AIT	0.41779	0.15368	0.01514	0.15455	0.16687	0.03166	0.07079	0.12248	0.02886	0.00188	**0.00147**
BR	0.11741	0.17177	0.02371	0.11645	0.11665	0.00485	0.07226	0.13313	0.00770	0.00029	**0.00019**
KP1123	0.25544	0.22335	0.01545	0.21989	0.24685	0.01399	0.05387	0.13585	0.01355	0.00058	**0.00024**
KP1654	0.36066	0.19607	0.01888	0.27621	0.23131	0.01560	0.06410	0.12659	0.05044	0.00035	**0.00030**
LE	0.18971	0.20332	0.02036	0.16489	0.14370	0.00796	0.06897	0.14908	0.01751	0.00082	**0.00033**
MID	0.21509	0.20378	0.02207	0.18090	0.17750	0.00971	0.07838	0.13583	0.02381	0.00060	**0.00028**
ME	0.52783	0.30672	0.01936	0.33994	0.32261	0.02138	0.06661	0.15255	0.03734	**0.00090**	0.00106
MV1	0.52887	0.31160	0.01561	0.21055	0.30008	0.01364	0.06330	0.16254	0.01785	0.00122	**0.00082**
MV18	0.26649	0.25937	0.01499	0.20357	0.25116	0.00715	0.05345	0.16153	0.01155	0.00023	**0.00022**
SP	0.12582	0.16205	0.01379	0.16681	0.14794	0.00682	0.04653	0.11416	0.00997	**0.00002**	**0.00002**
SBS	0.25892	0.18401	0.02574	0.32388	0.24738	0.00913	0.08103	0.13493	**0.00000**	0.00203	0.00147
SIT	0.18091	0.25074	0.02745	0.15135	0.17251	0.00783	0.09139	0.13453	0.01798	0.00171	**0.00124**

Table 3.4: Quantitative comparisons of average structural similarity index

Methods / Images	CBF	WLS	CSR	RP	RP-SR	LatLRR	CS-MCA	CNN	DNN	DL	Proposed
CM	0.62376	0.72827	0.74954	0.68137	0.66828	0.76182	0.72064	0.70593	0.70536	0.77758	**0.77773**
ST	0.49861	0.66873	0.67474	0.56186	0.49782	0.67029	0.66204	0.64162	0.64394	**0.68125**	0.68042
MFT	0.61724	0.72050	0.76172	0.70548	0.69924	0.76892	0.73225	0.72259	0.71112	0.78623	**0.78638**
AIT	0.67194	0.84142	0.86480	0.83519	0.81738	0.86276	0.84326	0.82597	0.82501	0.87467	**0.87486**
BR	0.61793	0.64769	0.66693	0.63219	0.62424	0.68909	0.63460	0.62636	0.62314	**0.70616**	0.70612
KP1123	0.64975	0.72693	0.76111	0.67889	0.65247	0.77222	0.74062	0.72790	0.71975	0.78692	**0.78730**
KP1654	0.53699	0.69730	0.71927	0.62293	0.63508	0.73233	0.69720	0.68246	0.66818	0.74651	**0.74655**
LE	0.69888	0.74456	0.78059	0.73414	0.72757	0.79005	0.74930	0.73288	0.73150	0.80594	**0.80621**
MID	0.59021	0.68192	0.71294	0.65640	0.65196	0.72292	0.68254	0.67433	0.66418	0.73985	**0.73997**
ME	0.45747	0.66894	0.72178	0.60058	0.59287	0.72532	0.70375	0.67747	0.68157	0.73502	**0.73529**
MV1	0.50982	0.72919	0.77048	0.69698	0.63134	0.77284	0.75505	0.72093	0.71527	0.78256	**0.78295**
MV18	0.68824	0.76997	0.81841	0.75496	0.72086	0.83420	0.79671	0.77181	0.77540	0.84742	**0.84772**
SP	0.63683	0.67587	0.68663	0.62227	0.62255	0.71667	0.65416	0.65309	0.64363	0.73217	**0.73247**
SBS	0.53005	0.67908	0.70304	0.59670	0.62040	0.71191	0.67525	0.65557	0.65703	**0.72860**	0.72849
SIT	0.68207	0.73718	0.77933	0.74709	0.72367	0.79277	0.75171	0.73514	0.72770	**0.80901**	0.80866

Table 3.5: Quantitative comparisons of average edge preservation index

Methods / Images	CBF	WLS	CSR	RP	RP-SR	LatLRR	CS-MCA	CNN	DNN	DL	Proposed
CM	0.68093	0.74477	0.77789	0.69472	0.69107	0.81879	0.76101	0.74752	0.76442	0.83152	**0.83204**
ST	0.60258	0.66523	0.69587	0.52280	0.48176	0.71043	0.69155	0.69811	0.63488	0.72348	**0.72671**
MFT	0.38920	0.54510	0.59472	0.55984	0.55803	0.68432	0.55419	0.54445	0.57417	0.70270	**0.70955**
AIT	0.77945	0.92920	0.94178	0.92362	0.90167	0.94880	0.93411	0.92374	0.94051	0.95315	**0.95363**
BR	0.58591	0.60933	0.66039	0.62096	0.62976	0.71873	0.63206	0.61184	0.61871	0.74439	**0.74794**
KP1123	0.59079	0.71674	0.72890	0.61052	0.58267	0.77093	0.71778	0.71832	0.72229	0.79187	**0.79382**
KP1654	0.57477	0.69082	0.71296	0.59284	0.58582	0.73989	0.69871	0.65968	0.70428	0.77252	**0.77309**
LE	0.42554	0.49105	0.54321	0.53619	0.52919	0.63076	0.50955	0.50200	0.52519	0.65282	**0.66287**
MID	0.39622	0.49706	0.53012	0.53254	0.53038	0.63693	0.48716	0.47987	0.47221	0.65574	**0.66781**
ME	0.45498	0.69038	0.74923	0.51199	0.54174	0.78169	0.73229	0.73960	0.70806	0.80071	**0.80128**
MV1	0.54331	0.73843	0.75911	0.69273	0.65109	0.80500	0.74833	0.74562	0.76353	0.81527	**0.81836**
MV18	0.68496	0.79267	0.81168	0.74566	0.71477	0.84746	0.80166	0.80356	0.79386	0.86333	**0.86362**
SP	0.64656	0.67145	0.68749	0.60725	0.60090	0.75524	0.65481	0.65949	0.66299	0.78255	**0.78303**
SBS	0.31405	0.51057	0.53468	0.50252	0.48912	0.61281	0.51688	0.48557	0.46178	0.63151	**0.64413**
SIT	0.45947	0.52624	0.55940	0.59872	0.59527	0.64858	0.53314	0.51942	0.51058	0.66813	**0.68093**

Table 3.6: Quantitative comparisons of average values of the FMI_{dct}, N_{abf}, $SSIM_a$ and EPI_a on TNO database

Evaluation measures /Methods	Avg.FMI_{dct}	Avg.N_{abf}	Avg.$SSIM_a$	Avg.EPI_a
CBF [116]	0.26309	0.31727	0.59957	0.57240
WLS [117]	0.33102	0.21257	0.72360	0.67837
CSR [34]	0.34640	0.01958	0.75335	0.71130
RP [24]	0.28210	0.22677	0.68424	0.64488
RP-SR [24]	0.27930	0.21444	0.67385	0.63737
LatLRR [129]	0.33817	0.01596	0.76486	0.76223
CS-MCA [35]	0.35841	0.06680	0.72964	0.69154
CNN [36]	0.35269	0.13280	0.71372	0.68444
DNN [37]	0.36658	0.02324	0.70852	0.68552
DL [40]	0.40463	0.00120	0.77803	0.77923
Proposed	**0.40597**	**0.00085**	**0.77831**	**0.78359**

3.3.5 Discussions and Future Works

The good quality of the image of a scene is unable to capture by the sensors due to the uncertainty in the source images and sensor noise. In this context, image fusion plays an essential role. The aim of the fusion task is to generate a fused image that incorporates the complementary data conveyed by various source images: thermal and visible images. However, generating a fused image with significant details and reduced artifacts is challenging in the vision-based system. In this chapter, we have developed two feature level IR and visible image fusion schemes: integration of bi-dimensional empirical mode

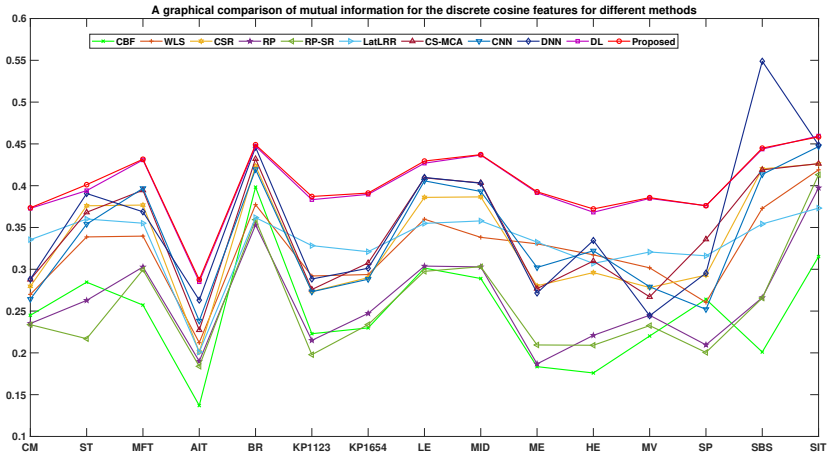


Figure 3.9: Quantitative comparisons of mutual information for the discrete cosine features for different schemes.

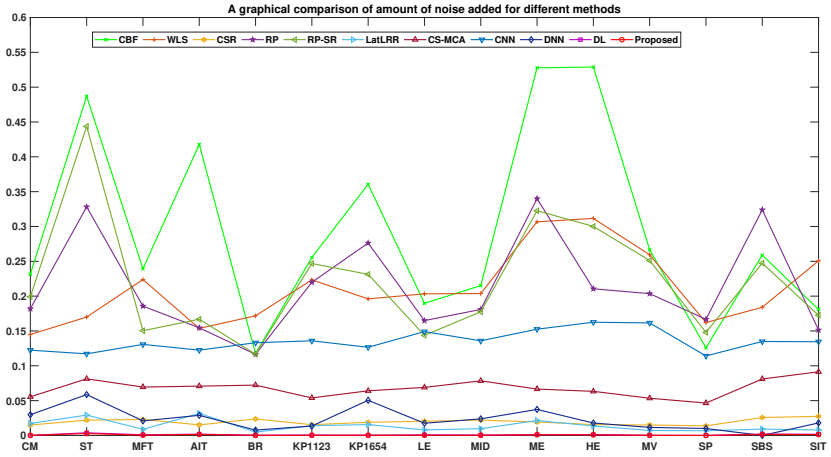


Figure 3.10: Quantitative comparisons of amount of noise added for different schemes.

decomposition with two streams VGG-16, and non-subsampled contourlet transform induced two streams ResNet-50 network. The proposed algorithms results are validated qualitatively as well as quantitatively by comparing with its result those of the different state-of-the-art (SOTA) techniques. For fair evaluation, the SOTA techniques are implemented without altering the parameters. It may be found that the proposed algorithms are attained better accuracy against several SOTA techniques.

Also, to know the achievement of the proposed algorithms, we have performed a quantitative comparison among integration of bi-dimensional empirical mode decomposition with two streams VGG-16 technique and non-subsampled contourlet transform induced

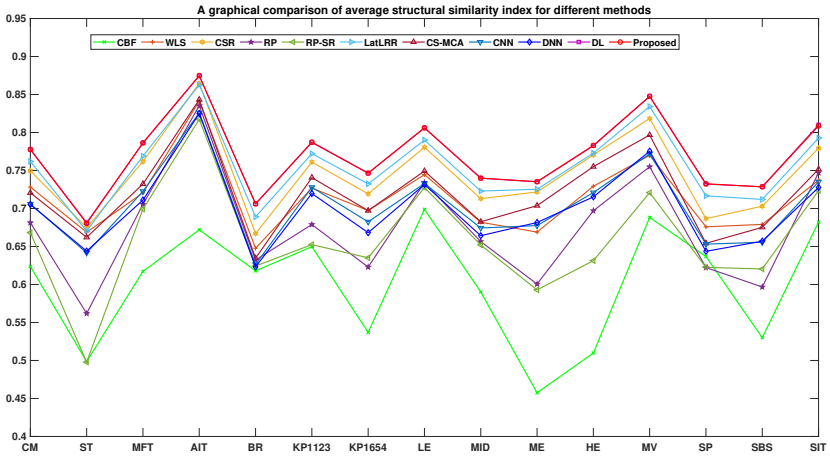


Figure 3.11: Quantitative comparisons of average structural similarity for different schemes.

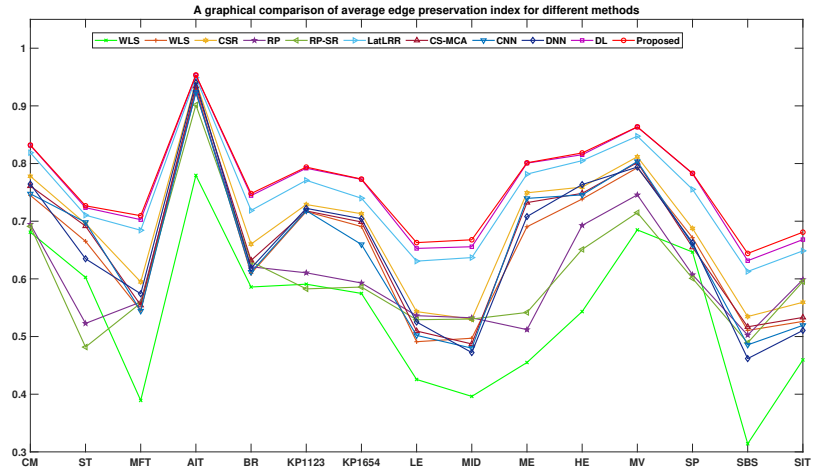


Figure 3.12: Quantitative comparisons of edge preservation index for different schemes.

two streams ResNet-50 network technique. From Table 3.7, it may be observed that the proposed non-subsampled contourlet transform induced two streams ResNet-50 network algorithm introduces much lesser noise and artifacts with maximum details into the fused image against the other proposed feature level infrared and visible image fusion scheme.

The proposed integration of bi-dimensional empirical mode decomposition with two streams VGG-16 and non-subsampled contourlet transform induced two streams ResNet-50 network schemes enhance visual perception of the thermal sequences with reduced artifacts. However, in the proposed integration of bi-dimensional empirical mode decomposition with two streams VGG-16 technique, the BEMD mechanism introduces artifacts in the fused image. Thus, we are planning to exploit the probabilistic-based decomposi-

Table 3.7: Quantitative comparisons between the proposed integration of bi-dimensional empirical mode decomposition with two streams VGG-16 and non-subsampled contourlet transform induced two streams ResNet-50 network schemes

Algorithms/ Quantitative Measures	integration of BEMD with two streams VGG-16	NSCT induced two streams ResNet-50 network
Avg.FMI_{dct}	0.39962	**0.40597**
Avg.N_{abf}	0.00149	**0.00085**
Avg.$SSIM_a$	0.77671	**0.77831**
Avg.EPI_a	0.77909	**0.78359**

tion strategy to handle the said issue in the future. Again, in the non-subsampled contourlet transform induced two streams ResNet-50 network algorithm, we have considered a deterministic weight map generation process that reduces the fused image's contrast. Considering uncertainty within a pixel in an image, a fuzzy set-theoretic fusion strategy can also be used.

3.4 Conclusions

Two image fusion schemes of IR and visible images at the feature level have been addressed in this chapter. In the proposed integration of bi-dimensional empirical mode decomposition with two streams VGG-16 technique, the proposed bi-dimensional empirical mode decomposition (BEMD) strategy is integrated with a VGG-16 deep neural architecture that can learn a mapping from image space to feature space at multi-scale with different levels. The proposed multi-level fusion strategy; investigates the spatial inter-dependency among these features and accurately acquires the complementary information from the source images. The proposed technique produces a fused image with essential details and reduced artifacts for IR and visible image pairs. Again, in the proposed non-subsampled contourlet transform induced two streams ResNet-50 network scheme, the proposed non-subsampled contourlet transform (NSCT) induced two-stream network using ResNet-50 architecture can exploit the multi-scale, multi-directional, and shift-invariant details of the sources at low-frequency and high-frequency bands. We have adhered to a pre-trained ResNet-50 deep neural network to generate the deep feature maps of the directional details. The deep layered architecture of the two-stream ResNet-50 network reduces the information loss during feature extraction and provides better accuracy as it is based on residual connections with identity mapping. We have proposed a unique fusion strategy

for deep multi-layer, shift-invariant features to detect the complementary information of source images efficiently. The proposed scheme produces a fused image with lesser noise for the corresponding IR and visible image pairs.

The results obtained by the proposed schemes are verified on various challenging scenes: illumination variation, smoke, occluded objects, non-uniform lighting conditions, etc. available at the *TNO* benchmark database. The efficacy of the proposed integration of bi-dimensional empirical mode decomposition with two streams VGG-16 scheme is corroborated against fifteen existing fusion schemes. Also, the performance of the proposed non-subsampled contourlet transform induced two streams ResNet-50 network algorithm is demonstrated against ten existing fusion schemes. To confirm our findings, we have used qualitative and quantitative analysis. It is observed that the fused images attained by the proposed algorithms have a strong correlation with the source images and higher accuracy than the existing fusion methods.

Chapter 4

Kernel Induced Possibilistic Fuzzy Associate Background Subtraction for Moving Object Detection

4.1 Introduction

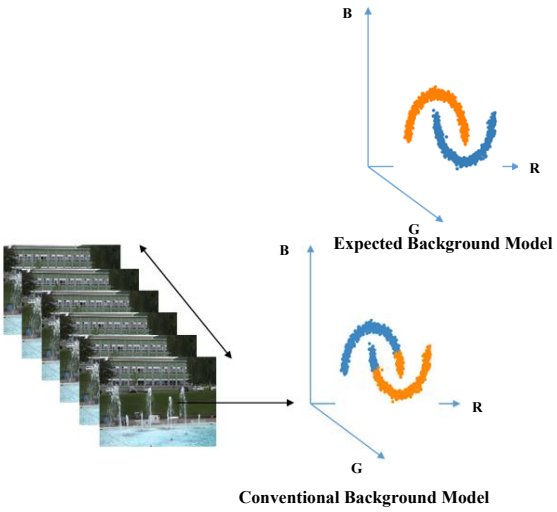


Figure 4.1: Data distribution due to conventional modeling

It may be summarized from the Chapter 1 that many works were reported in the state-of-the-art literature for background subtraction or local change detection. However, there are two major disadvantages that affect the accuracy and complexity of the algorithms: estimation of the parameters in the generative model used for background construction and unnecessary inclusion of outliers (non-required/noisy pixel values) in the background

model. The parameters used in the generative models in background construction are very complex to be calculated. In this regard, non-parametric methods are convenient to characterize the background of a scene as do not assume a generative model for describing the data. However, the use of multi-valued background models is always unable to characterize the randomness within the pixel values. Similarly, during the background construction process, the use of noisy or outliers pixels may produce errors in constructing a stable background model. The contribution of the noisy pixel may unnecessarily either increase the variance of a particular background type as the mode of the distribution may deviate or will try to merge multiple background types into a single one. Let's consider an example of a sequence of frames from *Fountain-02*, whose pictorial view of the same can be seen in Figure 4.1. Here different background types are represented by different colors. Considering the ideal shape of the data distribution one can expect that a good background construction algorithm will project the data distribution to follow the actual shape as shown at top of the figure. However, the shape of the pixel distributions using the conventional background model will project them to the group in the wrong background type, due to the varying density of the data point. Similar observations can be made for different other data distributions as shown in Figure 4.2. Further as can be seen in the last plot of Figure 4.2, due to noise points the background mode is shifted and produce a biased distribution, where the noise points try to shift the mode of the distribution.

Most of the techniques developed in the literature use deterministic approaches to resolve the problem of background construction. However, it is to be noted that the changes in subsequent frames of a video are quite common and the region undergoes static-background changes (pixel posses multi-valued background) are obvious. This high ambiguity in the spatial and temporal domain in the video is due to the multi-valued brightness of pixels. Deterministic approaches are effective only for video frames with significant contrast changes. Deterministic approaches are rarely able to distinguish the randomness of changes in video frames. Hence deterministic approaches may not always be suitable to produce good results in background subtraction. As reported in Artificial intelligence techniques, there are two approaches to deal with randomness: probabilistic and fuzzy set-theoretic approaches. However probabilistic theory-based BGS make a hard decision for foreground and background and involve many complex parametric estimation processes.

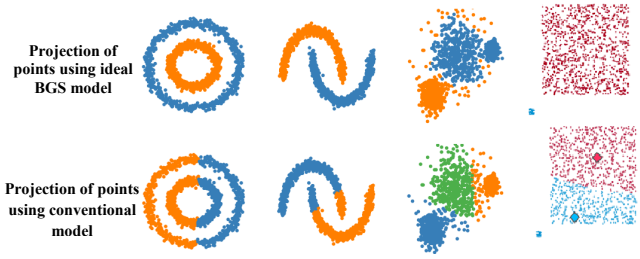


Figure 4.2: Conventional against ideal BGS model

Considering the above discussions, it can be understood that, fuzzy set theories are repute to handle uncertainties to a reasonable extent, arising from deficiencies of information available from a situation (the deficiency may result from incomplete, ill-defined, not fully reliable, vague, and contradictory information). It justifies applying the concept of fuzzy set based BGS is better than a hard decision-based BGS. In this context, it is also important that learning algorithms may not always able to provide useful insights structures of the data in the temporal domain which can help in the process of decision making/classifying a pixel in foreground/background class. A drawback with fuzzy set-theoretic BGS is its poor performance against noisy or outlier data and its accuracy degrades with poor or wrong initialization of background pixels. Again, the non-linear temporal ambiguities of data in lower dimensional space are also not able to give good results. This motivates us to design a fuzzy Possibilistic background subtraction algorithm in Kernel induced space is expected to yield a satisfactory performance in this regard. The use of possibilistic concepts in BGS will help in detecting and avoiding noisy/outlier points in background subtraction. Similarly, inducing in kernel space will help in mapping the spatiotemporal video data to a non-linear high dimensional kernel space of infinite dimensions, where it will be easy to build a stable background model.

In this chapter, we proposed a kernel induced possibilistic fuzzy associated unsupervised background subtraction technique to detect the local changes in fixed camera captured sequences. The proposed scheme follows two stages: background training and foreground segmentation. In the background construction stage, each pixel is modeled using a possibilistic fuzzy cost function in kernel induced space. The use of induced kernel function projects the low dimensional data into a higher dimensional feature space and the use of possibilistic function will construct a robust background model based on the density of the data in temporal direction avoiding the noisy and outlier points. Hence, the use of a possibilistic induced kernelized fuzzy modal variation cost function reduces

the effects of high ambiguity in the spatial and temporal domain of the video due to the multi-valued brightness of the pixels.

The performance of the proposed kernel induced possibilistic fuzzy associate background subtraction scheme is tested on the database: *changedetection.net*. The efficacy of the proposed scheme is evaluated on different performance evaluation measures. The investigation is corroborated by comparing the results against twenty-nine existing state-of-the-art techniques and is found to be better.

The rest of this chapter is organized as follows. The proposed kernel induced possibilistic fuzzy associate background subtraction for video scene is discussed in Section 4.2. Section 4.3 discusses the results and discussions with future works. The conclusion of the proposed works is carried out in Section 4.4.

4.2 Proposed Kernel Induced Possibilistic Fuzzy Associate Background Subtraction for Video Scene

A kernel induced possibilistic fuzzy associated background subtraction scheme is proposed in this chapter to detect the local changes corresponding to dynamic changes in the scene. The use of the induced kernel function will project the temporal data to an infinite-dimensional space to build the background model. To boost the accuracy against the noise and reduce the error due to outliers, the concept of possibilistic fuzzy C-means algorithm [133] is adhered to in the proposed background subtraction scheme. The flow-chart of the proposed scheme is provided in Figure 4.3.

Here it is assumed that the time instant is the same as that of the frame instant. Let us assume there are N frames are there in the video. The proposed scheme is divided into two stages: training or background construction and foreground separation. In the proposed scheme we assume n is the number of frames used for background construction or training and $N - n$ are the testing or target frames used for foreground separation. Each frame in the video is assumed to be represented as $I_i(x, y)$, where (x, y) represents the pixel location in a video frame and i is the frame instant, where $i = 0, 1, 2, .., n, ..., N$.

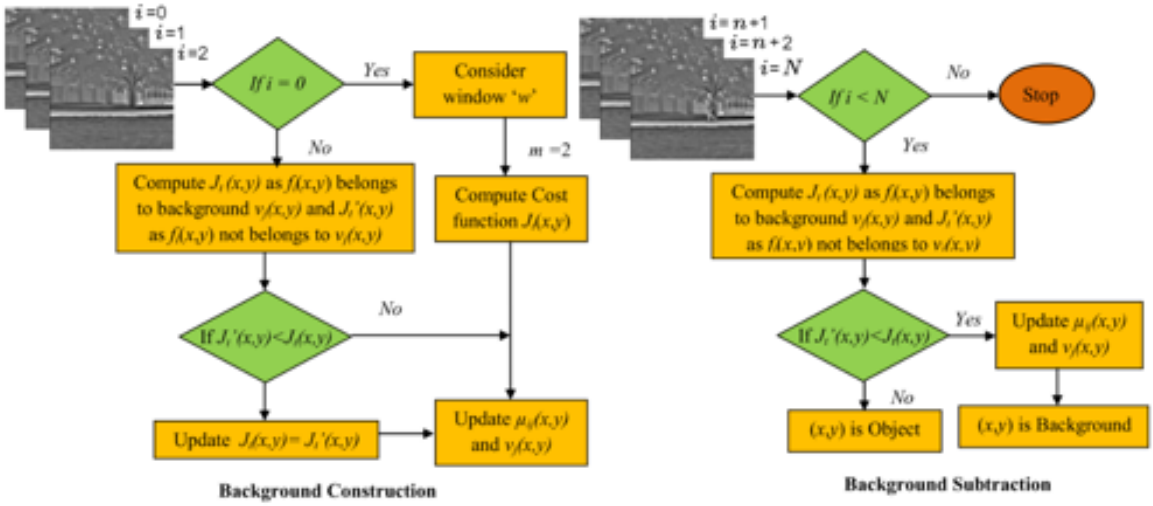


Figure 4.3: Block diagram of the proposed kernel induced possibilistic fuzzy associate background subtraction scheme.

4.2.1 Background Construction

In the proposed scheme we have used the n number of initial frames of a video to model the background of the scene. The model is integrated at individual pixel location (x, y). The proposed background model is initialized by considering a small region of support at every pixel location of the frame at $t = 0$ time instant. It is assumed that there are initially, two background types are there and further new background types are added based on the scene under consideration or sequence of image frames under consideration. Each background type is represented by the mode corresponding to it. In the next frame onward, each pixel at location (x, y) is fitted with the kernel induced possibilistic fuzzy associated cost function $Q_t(x, y)$ as follow,

$$Q_t(x,y) = \sum_{i=1}^{t}\sum_{j=1}^{m}\{\mu_{ij}^{r}(x,y)\}||\Phi(I_i(x,y)) - \Phi(v_j(x,y))||^2 + \sum_{j=1}^{m}\gamma_j\sum_{i=1}^{t}(\mu_{ij}(x,y)ln\mu_{ij}(x,y) - \mu_{ij}(x,y)), \quad t \leq n, \tag{4.1}$$

where m represents the number of constructed background type at location (x, y). The kernel function is assumed to be Φ which projects the pixel values from a RGB color plane to infinite dimensional plane and r is the fuzzification parameter, which induces degree fuzziness to the function. $\mu_{ij}(x, y)$ represents the belongingness of $I_i(x, y)$ pixel into j^{th} background type and v_j is defined as the mode corresponding to j^{th} background type. γ_j represents the spread of the j^{th} background type. The function ln in eq (4.1)

represents the logarithmic operation. The above mentioned cost function has two parts: the first part represents the within background type variance or average (typical) soft induced distortion $Q_{(t,KF)}(x,y)$ and the second part is the integration of the possibilistic term $Q_{(t,PM)}(x,y)$. Hence, this can be expressed as,

$$Q_t(x,y) = Q_{(t,KF)}(x,y) + Q_{(t,PM)}(x,y). \tag{4.2}$$

In eq (4.1), the term $||\Phi(I_i(x,y)) - \Phi(v_j(x,y))||^2$ is the distance in the kernel space. Using Mercer's theorem the distance in kernel space can be computed as,

$$||\Phi(I_i(x,y)) - \Phi(v_j(x,y))||^2 = K(I_i(x,y), I_i(x,y)) + K(v_j(x,y), v_j(x,y)) - 2K(I_i(x,y), v_j(x,y)). \tag{4.3}$$

where the $K(.)$ represents the dot product and considering a positive semidefinite kernel for any two entity x_1 and x_2 we may describe it as;

$$K(x_1, x_2) = \Phi(x_1)^T \Phi(x_2). \tag{4.4}$$

Considering a Gaussian kernel function we can express;

$$K(I_i(x,y), v_j(x,y)) = exp\{-||I_i(x,y) - v_j(x,y)||^2/\sigma^2\}, \tag{4.5}$$

where σ is the variance corresponding to each background type. Hence the expression for the cost function can be given as,

$$\begin{aligned} Q_t(x,y) &= \sum_{i=1}^{t}\sum_{j=1}^{m} \mu_{ij}^r(x,y)(1 - K(I_i(x,y), v_j(x,y))) \\ &+ \sum_{j=1}^{m} \gamma_j \sum_{i=1}^{t} (\mu_{ij}(x,y) ln \mu_{ij}(x,y) - \mu_{ij}(x,y)). \quad t \leq n. \end{aligned} \tag{4.6}$$

At t^{th} time instant for pixel location (x,y), the value $I_t(x,y)$ is fitted to the cost function in eq (4.6) once by computing the cost function considering the pixel belongs to the mode of m existing background type and another by considering the value $I_i(x,y)$ is added as new background type (hence total $m+1$ background type), where new mode is $I_t(x,y)$. Then, the minimum cost value along with its parameters will be used to update the old

cost value. Considering this, the mode value, fuzzy membership value are updated for each pixel value in different frames of the video.

Considering $r = 1$ and taking the derivatives of the above mentioned cost function eq (4.6), we may obtain the updating rule for the fuzzy membership function as;

$$\mu_{ij}(x,y)^{t+1} = e^{-\left\{\frac{2(1-K(I_t(x,y),v_j(x,y))}{\gamma_j}\right\}}, \tag{4.7}$$

similarly, the mode corresponding to the background type in kernel space can be obtained as,

$$v_j(x,y)^{t+1} = \frac{\sum_{i=1}^{t} \mu_{ij}(x,y)\phi(I_i(x,y))}{\sum_{i=1}^{t} \mu_{ij}(x,y)}. \tag{4.8}$$

Hence for each pixel location in $(t+1)^{th}$ frame the updated background model's fuzzy membership value and modes can be represented as,

$$v_j(x,y)^{t+1} = \begin{cases} \frac{\sum_{i=1}^{t} \mu_{ij}(x,y)\phi(I_i(x,y))}{\sum_{i=1}^{t} \mu_{ij}(x,y)}, \\ \quad \text{if } I_i(x,y) \in \text{old background;} \\ \text{initiate } v_{m+1}(x,y), \text{ otherwise} \end{cases}. \tag{4.9}$$

The updated membership values will be,

$$\mu_{ij}(x,y)^{t+1} = \begin{cases} e^{-\left\{\frac{2(1-K(I_t(x,y),v_j(x,y))}{\gamma_j}\right\}}, \ j = 1,2,3,..m \\ \text{if } I_i(x,y) \in \text{old background types;} \\ e^{-\left\{\frac{2(1-K(I_t(x,y),v_j(x,y))}{\gamma_j}, \ j=1,2,3,..,m,m+1\right\}} \\ \text{otherwise} \end{cases}. \tag{4.10}$$

However the parameters in eqs (4.7) and (4.10) can not be computed directly. It is to be noted that the induced mapping will map the data to infinite feature space by;

$$\|\phi(I_i(x,y))\| = \phi(I_i(x,y))^T \phi(I_i(x,y)) = k_{ii}. \tag{4.11}$$

Considering the above kernel tricks we may obtain the update rule for the membership function as,

$$\mu_{ij}(x,y) = exp\left[-\frac{1}{\gamma_j}\left(kii - \frac{2\sum_{l=1}^{t}\mu_{lj}(x,y))k_{lj}}{\sum_{i=1}^{t}\mu_{ij}(x,y))} + \frac{\sum_{l=1}^{t}\sum_{m=1}^{t}\mu_{lj}(x,y))\mu_{mj}(x,y))k_{lm}}{\left(\sum_{i=1}^{t}\mu_{ij}(x,y))\right)^2}\right)\right] \tag{4.12}$$

where $k_{lj} = K(I_l(x,y), I_j(x,y))$. Taking the partial derivative of the cost function in eq (4.1) with respect to γ_j and equating to 0, we may obtain,

$$\gamma_j^{t+1} = \lambda\{\frac{\sum_{i=1}^{t}\mu_{ij}(x,y)\|\Phi(I_i(x,y)) - \Phi(V_j(x,y))\|^2}{\sum_{i=1}^{t}\mu_{ij}(x,y)}\} \tag{4.13}$$

Deriving the above equation we may obtain,

$$\gamma_j^{t+1} = \frac{\lambda}{\sum\limits_{i=1}^{t}\mu_{ij}(x,y)}\left[\sum_{i=1}^{t}\mu_{ij}(x,y)\left(k_{ii} - \frac{2\sum\limits_{l=1}^{t}\mu_{lj}(x,y)k_{lj}}{\sum\limits_{i=1}^{t}\mu_{ij}(x,y)} + \frac{\sum\limits_{l=1}^{t}\sum\limits_{m=1}^{t}\mu_{lj}(x,y)\mu_{lj}(x,y)k_{lm}}{\left(\sum\limits_{i=1}^{t}\mu_{ij}(x,y)\right)^2}\right)\right], \tag{4.14}$$

where λ represents the spread function constant and is used for the controlling the spread of the modes corresponding to the background. The said steps will be followed for each pixel location for all n frames.

4.2.2 Foreground Separation and Background Update

In the next stage, of processing, the locations of the moving objects are detected from the test/target frames. The foreground segmentation or moving object detection is followed on $N - n$ frames starting from $(n+1)^{th}$ frame to N^{th} frame. At each pixel location in the target frame it is checked if the pixel belongs to any of the existing background mode type v_j by fitting it against the Kernel induced possibilistic function as in eq (4.6). If any pixel in the target frame belongs to any of the mode type v_j, then the pixel in the target frame will be considered as the part of the background else assumed as a foreground pixel. This can be represented as,

$$\psi_i(x,y) = \begin{cases} 0, \text{ if } I_i(x,y) \in v_j(x,y), \ i = n+1, ..., N; \\ 1, \text{ otherwise.} \end{cases} \tag{4.15}$$

$\psi_i(x,y)$ is local change detection output image, where the foreground pixels are represented by 1 and background by 0. $v_j(x,y)$ represents jth background modes at (x,y) location. Hence the background at each pixel location in target frame are updated as,

$$v_j(x,y) = \begin{cases} \frac{\sum_{i=1}^{t} \mu_{ij}(x,y)\phi(I_i(x,y))}{\sum_{i=1}^{t} \mu_{ij}(x,y)}, \text{ if } \psi_i(x,y) = 0; \\ v_j(x,y), \text{ otherwise} \end{cases} . \tag{4.16}$$

and the membership values are updated as;

$$\mu_{ij}(x,y) = \begin{cases} e^{-\left\{\frac{2(1-K(I_t(x,y),v_j(x,y))}{\gamma_i^t}\right\}} \text{ if } \psi_i(x,y) = 0; \\ \mu_{ij}(x,y), \text{ otherwise} \end{cases} . \tag{4.17}$$

and the new cost function is calculated as;

$$Q_t(x,y) = \begin{cases} \sum_{i=1}^{t}\sum_{j=1}^{m} \mu_{ij}^r(x,y)(1 - K(f_i(x,y), v_j(x,y))) \\ + \sum_{j=1}^{m} \gamma_j \sum_{i=1}^{t} (\mu_{ij}(x,y) ln \mu_{ij}(x,y) \\ - \mu_{ij}(x,y)), \\ \text{if } \psi_i(x,y) = 0; \\ Q_t(x,y), \text{ otherwise.} \end{cases} \tag{4.18}$$

Steps of the proposed background subtraction technique are enumerated in Algorithm 1.

4.3 Results and Discussions

The simulation and experimentation of the proposed scheme is carried out in a *Core* i7^{TM}, 7th generation system with 16GB RAM, 16MB L2 cache. The proposed technique is implemented in C++ programming language with Ubuntu operating system. It is validated on different test sequences and in this chapter, it is reported on the popular benchmark databases: *changedetection.net* [114].

Algorithm 1 : Induced Possibilistic Kernelized average soft induced distortion based background subtraction scheme

Input: N number of video frames $I_1, I_1, .., I_N$.
Divide N frames into training (n) and testing $(N-n)$ frames.
Proposed scheme has two stages: background model and foreground separation & update.
Training Stage:
Initiate with $i = 0$;
− Consider a rectangle w at every pixel location (x, y) and initiate the background, assuming mode $(m = 2)$.
− Compute the cost function using eq (4.1).
− Compute the fuzzy membership values and modes using eqs (4.7)-(4.8)
−While $(i \leq n)$
{
−i=i+1;
−for each pixel (x, y);
$*$ compute the new cost function $J_t(x, y)$ assuming that $f_i(x, y)$ belongs to the old background
$*$ compute the new cost function $J_t^{'}(x, y)$ assuming that $f_i(x, y)$ belongs to new mode
$*$ if $(J_t^{'}(x, y) < J_t(x, y))$
−Update $J_t(x, y) = J_t^{'}(x, y)$
−Update the new modes and membership values as in eqs (4.9)-(4.12)
}
Object separation and background update:
Initiate with $i = n + 1$;
−While $(i \leq N)$
{
−i=i+1;
−for each pixel (x, y);
$*$ compute the new cost function $J_t(x, y)$ assuming that $f_i(x, y)$ belongs to the old background
$*$ Then consider the pixel as a part of background else as foreground
−Update the new modes and membership values based on eqs (4.16)-(4.18)
}

4.3.1 Performance Evaluation

Here we have gone for two ways of evaluating the proposed scheme: visually and quantitatively.

4.3.1.1 Visual Analysis of Results

The visual analysis of results are carried out using different challenging sequences: *PETS2006*, *Badminton*, *Water surface*, *Canoe*, *MSA*, *Waving tree*, and *Snowfall*. The results obtained by the proposed scheme and other state-of-the-art techniques are reported in Figure 4.4. All the considered original frames are shown in Figure 4.4 (a). Corresponding ground-truth images are provided in Figure 4.4 (b). The KDE-based BGS scheme's results are reported in Figure 4.4 (c) which produces many missed alarms. Figure 4.4 (d) displays the results on the considered frames by the BRPCA based BGS scheme. It is observed from this figure that in several places the BRPCA based scheme produces many missed alarms and false alarms for low-resolution sequences. Figure 4.4 (e) represents the results of the ViBe based BGS scheme, where poor quality results are captured. The results for the pROST based BGS scheme are reported in Figure 4.4 (f), where better results are obtained for different sequences except the snowfall. Similar analysis can be made for the DPGMM scheme (as shown in Figure 4.4 (g)). The results obtained by the feature bags technique are shown in Figure 4.4 (h). The results obtained by the proposed BGS scheme as shown in Figure 4.4 (i) is found to have provided better accuracy with fewer misclassification errors.

A similar analysis is presented in Figure 4.5 where the comparison of the proposed scheme against different deep learning techniques are provided. Figure 4.5 (c) and (d) represent the object detection results obtained with those of the DeepBS and Cascade CNN schemes. Both the schemes provided results where many parts of the moving objects are not detected properly. Figure 4.5 (e) and (f) shows the results of the BSUV_net and BSUV_net+semantic BGS techniques where many parts of the scene are falsely identified as the moving object. The results of the proposed scheme as provided in Figure 4.5 (g) are found to be visually more prominent as compared to all other considered techniques.

The results obtained by the proposed scheme and other state-of-the-art techniques for the thermal sequences available at *changedetection.net* are reported in Figure 4.6. All the considered original frames are shown in Figure 4.6 (a). Corresponding ground-truth

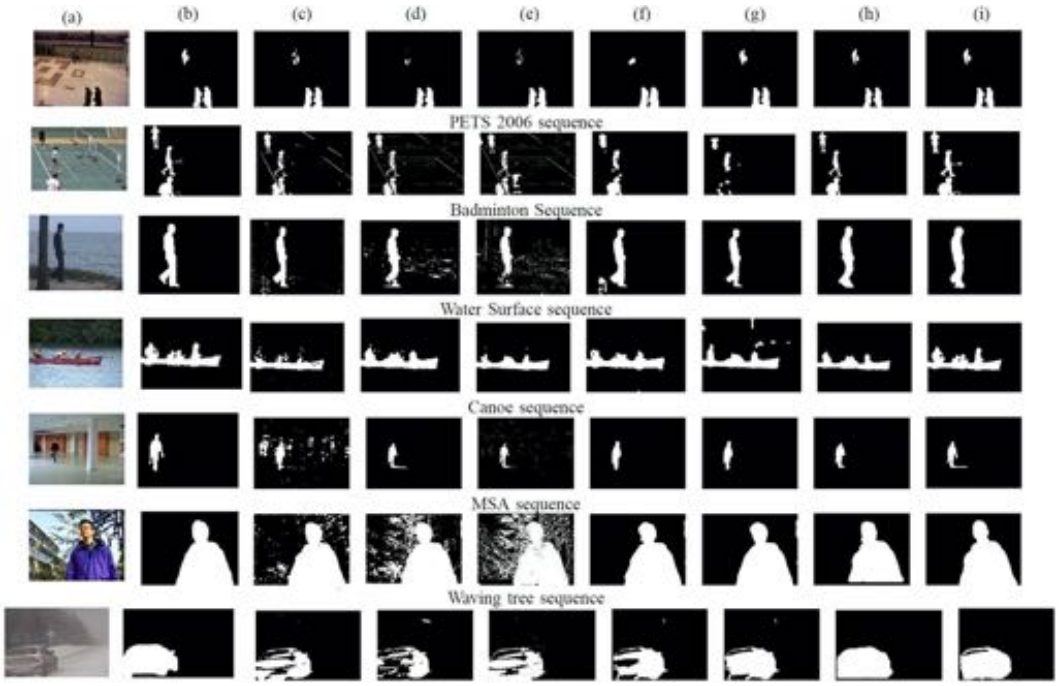


Figure 4.4: Moving object detection for different sequences (*PETS2006*, *Badminton*, *Water surface*, *Canoe*, *MSA*, *Waving tree*, and *Snowfall*): (a) original frame, (b) corresponding ground-truth, moving object detection results obtained by non-deep learning based BGS schemes: (c) KDE, (d) BRPCA, (e) ViBe, (f) pROST, (g) DPGMM, (h) feature bags and (i) proposed kernel induced possibilistic fuzzy associate background subtraction scheme.

images are provided in Figure 4.6 (b). The DeepBS based BGS scheme's results are reported in Figure 4.6 (c), which produces many missed alarms. Figure 4.6 (d) depicts the results obtained by the WisenetMD based BGS scheme. It is observed from this figure that the WisenetMD based scheme produces ghosts for the considered frames. Figure 4.6 (e) represents the results of the Cascade CNN based BGS scheme, where few pixels are misclassified as the object. The results for the IUTIS_5 based BGS scheme are reported in Figure 4.6 (f), where poor quality results are produced. Figure 4.6 (g), shows the results of the BSUV_net which produce a high false-positive rate. Figure 4.6 (h) denotes the results obtained by the SemanticBGS where many object pixels are wrongly classified as background pixels. The results of BSUV_net2.0 are presented in 4.6 (i), where false detections are observed. The results obtained by the proposed BGS scheme are shown in Figure 4.6 (j). It is found from this figure that the proposed algorithm precisely classifies the object as well as background pixels and provides a better accuracy.

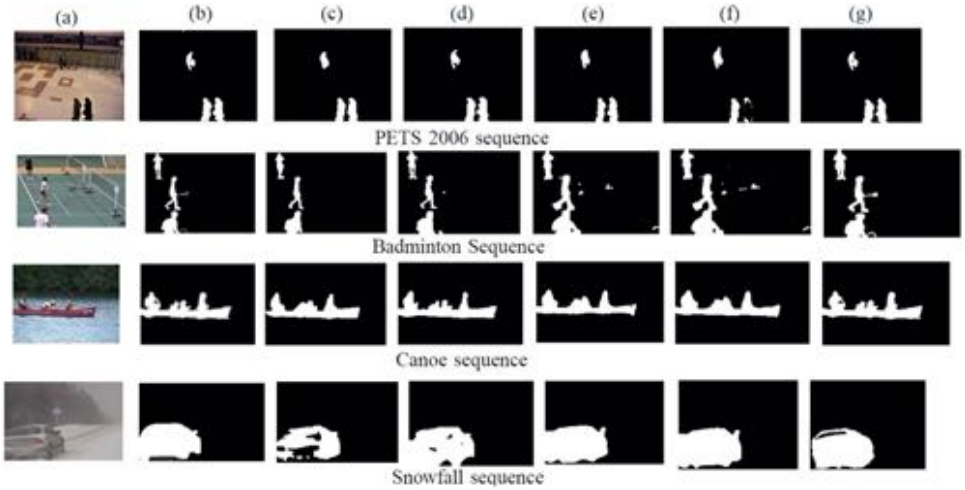


Figure 4.5: Moving object detection for different sequences: (a) original frame, (b) corresponding ground-truth, moving object detection results obtained by deep learning based BGS schemes: (c) DeepBS, (d) Cascade CNN, (e) BSUV_net, (f) BSUV_net+semantic, and (g) proposed kernel induced possibilistic fuzzy associate background subtraction scheme.

The execution time taken by the proposed scheme as compared to other considered existing techniques is reported in Table 4.1. It may be observed from this table that, the time taken by the proposed scheme per frame is more than ViBe and Kernelized fuzzy technique whereas lesser time as compared to the other considered techniques.

Table 4.1: Average execution time (in second) required for different algorithms

Video	Codebook	KDE	BRPCA	DT	ViBe	Gaussian Wronskian	pROST	DPGMM	Feature Bags	Fuzzy mode	Possibilistic Induced
Water surface	0.10	0.12	0.16	0.15	0.04	0.12	0.10	0.10	0.12	0.07	0.08
MSA	0.13	0.15	0.21	0.19	0.05	0.13	0.12	0.11	0.13	0.09	0.09
Waving tree	0.13	0.15	0.22	0.19	0.05	0.13	0.12	0.11	0.13	0.09	0.09
PETS2006	0.16	0.17	0.25	0.22	0.07	0.15	0.14	0.13	0.16	0.11	0.12
Badminton	0.14	0.16	0.21	0.19	0.06	0.12	0.12	0.12	0.13	0.11	0.11
Canoe	0.14	0.15	0.20	0.18	0.05	0.11	0.12	0.12	0.13	0.09	0.10
Snowfall	0.15	0.17	0.25	0.22	0.07	0.15	0.14	0.13	0.16	0.11	0.12

4.3.1.2 Quantitative Evaluation

The performance of the proposed scheme is evaluated by comparing it against the state-of-the-art BGS techniques: Codebook [67], KDE [56], BRPCA [79], DT [134], ViBe [70], Gaussian Wronskian [63], pROST [135], DPGMM [71], Feature Bags [73], Fuzzy mode[99], SOBS-CF [91], SuBSENSE [72], RPCA [82], multimode background [74], GMM [55], PAWCS [136], SharedModel [137], Spectral-360 [138], WeSamBE [75], Cascade CNN [103], DeepBS [101], BSUV_net [108], BSUV-net+SemanticBGS [108], BMN-BSN [139], BSPVGAN [112], WisenetMD [140], IUTIS-5 [141], SemanticBGS [142], and BSUV_net

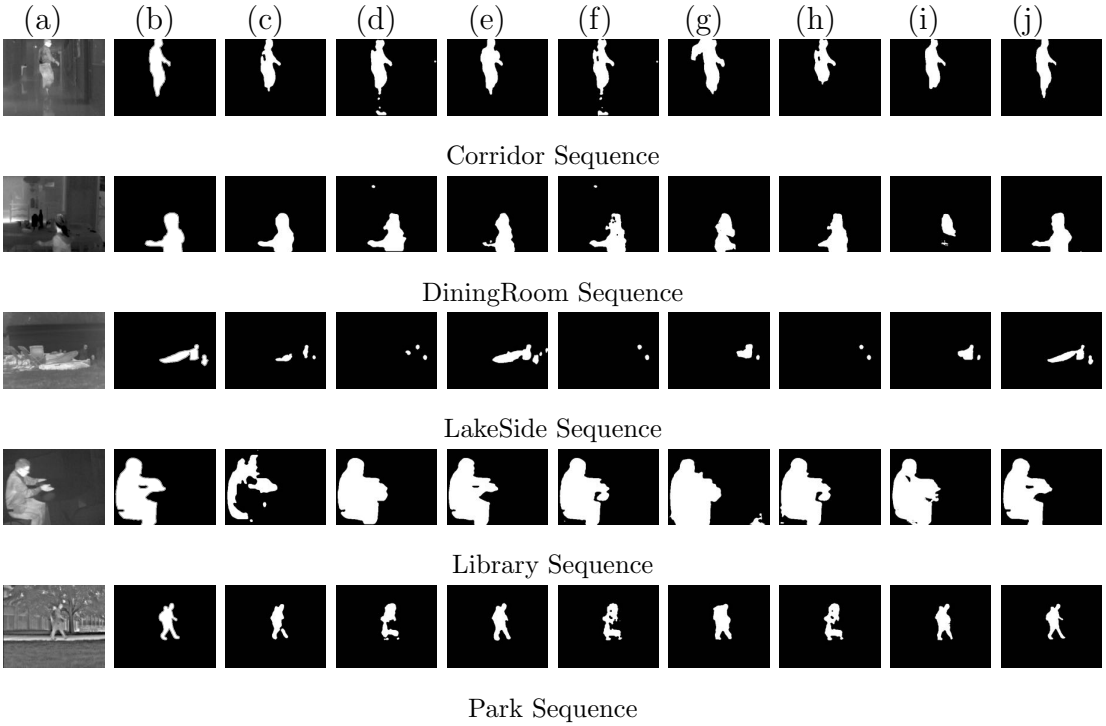


Figure 4.6: Moving object detection for different sequences: (a) original frame (b) corresponding ground-truth, moving object detection results obtained by deep learning based BGS schemes: (c) DeepBS, (d) WisenetMD, (e) Cascade CNN, (f) IUTIS_5, (g) BSUV_net (h) SemanticBGS, (i) BSUV_net2.0 and (j) proposed kernel induced possibilistic fuzzy associate background subtraction scheme.

2.0 [143].

The effectiveness of the proposed scheme on the *PETS2006*, *Badminton*, *Water surface*, *Canoe*, *MSA*, *Waving tree*, and *Snowfall* sequences are evaluated by using three evaluation measures: average Precision, average Recall, and average F-measure. Table 4.2 represents the evaluation of the proposed scheme on *PETS2006*, *Badminton*, *Water surface*, *Canoe*, *MSA*, *Waving tree*, and *Snowfall* sequences. In Table 4.2, we have compared the performance of the proposed scheme with those of the considered state-of-the-art techniques in terms of the average Precision, the average Recall, and the average F-measure. It may be observed that in the case of a waving tree sequence the proposed scheme is giving lesser Recall as compared to the feature bags scheme. Similarly, for the Canoe sequence, the Precision obtained by the KDE scheme is found to be higher than the proposed scheme. However, the proposed scheme provides a higher average F-measure in the case of all the sequences as compared to the considered state-of-the-art techniques.

The performance of the proposed scheme and other said schemes on *changedetection.net* database is provided in Table 4.3. It may be observed from this table that the

Table 4.2: Average Precision, Recall and F-measure for different image sequences

	Water surface			MSA			Waving tree			PETS2006			Badminton			Canoe			Snowfall		
Approaches	Pr	Re	FM	Pr	Re	FM	Pr	Re	FM	Pr	Re	FM	Pr	Re	FM	Pr	Re	FM	Pr	Re	FM
Codebook [67]	0.51	0.71	0.60	0.81	0.86	0.84	0.80	0.88	0.84	0.84	0.97	0.90	0.61	0.77	0.69	0.86	0.85	0.85	0.67	0.79	0.73
KDE [56]	0.40	0.79	0.52	0.67	0.79	0.73	0.53	0.81	0.64	0.83	0.79	0.81	0.67	0.79	0.73	**0.96**	0.83	0.89	0.90	0.68	0.78
BRPCA [79]	0.86	0.92	0.89	0.83	0.90	0.86	0.81	0.86	0.83	0.80	0.86	0.83	0.68	0.80	0.74	0.78	0.86	0.82	0.70	0.76	0.73
DT [134]	0.84	0.91	0.88	0.80	0.88	0.84	0.80	0.75	0.77	0.77	0.81	0.79	0.75	0.82	0.79	0.81	0.86	0.83	0.70	0.76	0.73
ViBe [70]	0.71	0.85	0.77	0.86	0.90	0.88	0.79	0.84	0.82	0.86	0.70	0.78	0.73	0.78	0.75	0.88	0.89	0.88	0.71	0.76	0.73
Gaussian Wronskian [63]	0.91	0.95	0.93	0.83	0.87	0.85	0.77	0.84	0.81	0.90	0.97	0.93	0.83	0.86	0.85	0.87	0.89	0.88	0.73	0.82	0.77
pROST [135]	0.65	0.78	0.72	0.85	0.93	0.89	0.79	0.88	0.83	0.70	0.67	0.68	0.87	0.81	0.84	0.92	0.93	0.92	0.66	0.76	0.71
DPGMM [71]	0.89	0.94	0.92	0.87	0.94	0.90	0.82	0.87	0.85	0.85	**0.98**	0.91	0.89	0.68	0.77	0.79	**0.95**	0.86	0.76	0.81	0.79
Feature Bags [73]	0.92	0.97	0.95	0.90	0.92	0.91	0.86	**0.96**	0.89	0.92	**0.98**	0.95	0.86	0.85	0.85	0.89	0.91	0.90	0.82	0.87	0.85
Possibilistic Induced	**0.95**	**0.98**	**0.96**	**0.94**	**0.95**	**0.94**	**0.90**	0.94	**0.92**	**0.94**	**0.98**	**0.96**	**0.92**	**0.93**	**0.92**	0.93	**0.95**	**0.93**	**0.90**	**0.94**	**0.92**

proposed scheme provides a better average F-measure output as compared to the existing considered sate-of-the-arts-techniques on as sequences of *changedetection.net* database. It may be observed that in two instances, dynamic background and camera jitter, the Cascade CNN scheme provided a better result. However, the proposed scheme provides comparably results in this regard. Also, the efficiency of the proposed scheme is corroborated against seventeen state-of-the-art BGS techniques on five thermal sequences available at *changedetection.net* database are provided in Table 4.4. It may be observed that the proposed scheme has provided a higher accuracy in terms of average Precision, average Recall, average F-measure, and lower values of the average PWC as compared to seventeen state-of-the-art techniques.

Table 4.3: Average F-measure for changedetection.net database

Techniques	Baseline	Dynamic Background	Camera Jitter	Intr. Obj. Motion	Shadow	Thermal
KDE [56]	0.909	0.596	0.572	0.409	0.803	0.742
SOBS-CF [91]	0.873	0.309	0.745	0.534	0.664	0.873
ViBe [70]	0.870	0.565	0.600	0.507	0.803	0.665
DPGMM [71]	0.929	0.814	0.748	0.542	0.813	0.813
SuBSENSE [72]	0.950	0.818	0.815	0.657	0.899	0.817
RPCA [82]	0.677	0.684	0.547	0.672	0.729	0.565
Feature Bags [73]	0.943	0.837	0.818	0.643	0.820	0.822
Multimode background [74]	0.932	0.621	0.836	**0.823**	0.838	0.910
WeSamBE [75]	0.936	0.790	0.780	0.724	0.914	0.813
Cascade CNN [103]	0.967	**0.947**	**0.967**	0.868	0.946	0.887
DeepBS [101]	0.965	0.844	0.896	0.689	0.943	0.650
BSUV_net [108]	0.969	0.797	0.774	0.750	0.922	0.858
BSUV_net+SemanticBGS [108]	0.964	0.818	0.779	0.760	**0.967**	0.845
BMN-BSN [139]	0.952	0.637	0.696	0.637	0.789	0.785
Possibilistic Induced	**0.972**	0.903	0.911	**0.823**	0.951	**0.913**

4.3.2 Discussions and Future Works

The proposed scheme is tested on different sequences with challenging background conditions: speckling water, vibrating blinds, vibrating trees, snowfall, water fountains, shadow,

Table 4.4: Quantitative comparisons on 5 thermal sequences of changedetection.net database

Approaches	Avg.Precision	Avg.Recall	Avg.F-Measure	Avg.PWC
KDE [56]	0.8974	0.6725	0.7423	1.6795
GMM [55]	0.8652	0.5691	0.6621	4.2642
PAWCS [136]	0.8280	0.8504	0.8324	1.4018
SuBSENSE [72]	0.8328	0.8161	0.8171	2.0125
SOBS-CF [91]	0.8715	0.6347	0.7140	1.8021
WeSamBE [75]	0.8554	0.7727	0.7962	2.3538
Multimode Background [74]	0.8268	0.8162	0.8194	1.4289
SharedModel [137]	0.8072	0.8618	0.8319	1.8656
Spectral-360 [138]	0.9114	0.7238	0.7764	1.6337
DeepBS [101]	0.9257	0.6637	0.7583	3.5773
BSPVGAN [112]	0.9770	0.9763	0.9764	0.2406
WisenetMD [140]	0.8696	0.7867	0.8152	1.8993
Cascade CNN [103]	0.8577	0.9461	0.8958	1.0478
IUTIS-5 [141]	0.8969	0.7990	0.8303	1.1484
BSUV_net [108]	0.8551	0.8739	0.8581	1.7058
SemanticBGS [142]	0.9118	0.7664	0.8219	1.3897
BSUV_net 2.0 [143]	0.9359	0.8594	0.8932	1.1659
Proposed	**0.9861**	**0.9838**	**0.9849**	**0.1251**

foggy scene, rainy scene, surface reflection, etc. The proposed scheme is tested on the scene with camera jitter, thermal captured sequences, camera shake and vibration, the scene with different illumination conditions, and real-life long run sequences to establish the effectiveness of the proposed scheme. It may be concluded that the proposed scheme is found to be very effective in constructing a stable and robust background model with the ability to deal with the randomness of pixels in video scenes due to multi-valued background brightness.

The inclusion of the fuzzy set theory in the proposed BGS technique involves a soft decision in infinite-dimensional kernel induced space to construct the background and detect the moving object locations in a video frame. It may also be noted that the use of the possibilistic concept in fuzzy set theory in BGS will help to handle the uncertainties in a video frame up to a reasonable extent, arising from deficiencies of information available from a situation. This deficiency may be due to incomplete, ill-defined, not fully reliable, vague, and contradictory information.

In the proposed scheme few important parameters are considered for the initialization of the experiments. We initialize our algorithm with a Gaussian kernel, with $\gamma = 1$. Further γ is updated based on eq (4.14). We also tried with random initialization of γ value in the range $[0.01, 10]$ and found that the performance is not significantly changing, hence fixed the initialization value $\gamma = 1$. The parameter λ is set to be 1 as optimum on a trial and error basis. We tried with different values of λ and found that performance is

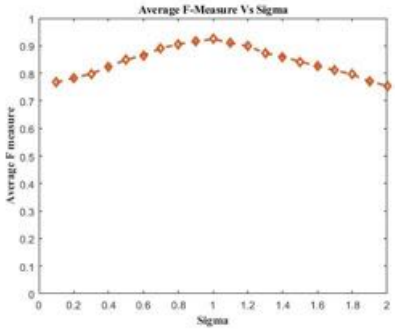


Figure 4.7: Selection of optimum value of parameter σ on changedetection.net database.

not significantly changing. We checked the Gaussian kernel with different values of σ in the range $[0.1,\ 2]$ with F-measure values on the Changedetection.net database. The same is reported in Figure 4.7. We found that with $\sigma = 1$, we are getting the best F-measure value and we adhered to it.

In many instances, the objects in the scene stay for some time and move further or vice-versa. In the proposed scheme we update the background model during the testing phase, we check if at instances a foreground/background doesn't change for at least 48 frames, then we start updating it as a new background mode type and further consider it as a part of the background model.

The proposed scheme is evaluated against considered state-of-the-art techniques. The parameters used for the state-of-the-art techniques are considered to be an optimum set of values. The codebook based BGS technique, considers two learning rate parameters in the range $[0.2, 0.7]$ and $[1.1, 1.8]$. The codebook based BGS's source code is obtained from [1]. For ViBe BGS the parameters are considered as follow: total samples = 30, matching threshold = 10, matching number = 2 and update-factor = 8. In Gaussian modeled Wronskian technique the threshold parameter is considered in the range $[3, 5]$ and the learning rate parameter in the range $[0.001, 0.3]$. The source code for the same is obtained from [2]. For pROST algorithm we have used $k = 15$, $p = 0.25$, and $mu = 0.025$. Similarly for DPGMM algorithm, we have used $min_{size} = 32$, $cert\ limit = 0.005$, $conc = 0.01$, $max\ layers = 8$, and $threshold = 0.6$.

The proposed scheme is eligible to detect the local changes from the different challenging scenarios. It may be observed that the fuzzy membership function is itself induces uncertainty. Hence this may give poor results sometimes. In such a scenario, the use

[1] http://www.umiacs.umd.edu/ knkim/
[2] https://sites.google.com/site/subudhibadri/mywork/bgs

of a type-II fuzzy set may produce an effective result by reducing the effects of uncertainty. Further in challenging sequences with non-static background conditions, it may be expected that the fuzzy histogram-based BGS technique may produce an improved accuracy.

4.4 Conclusions

In this chapter, we put forth, a background subtraction technique using kernel induced possibilistic fuzzy theoretic decision process. In the background construction stage, each pixel is modeled using a possibilistic fuzzy cost function in kernel induced space. The use of the induced kernel function will project the low dimensional data into a higher dimensional space and the use of the possibilistic function will construct a robust background model based on the density of the data in the temporal direction avoiding the noisy and outlier points. The performance of the proposed scheme is tested on the database: *changedetection.net*. The effectiveness of the proposed scheme is evaluated on different performance evaluation measures. The investigation is corroborated by comparing the results against twenty-nine existing state-of-the-art techniques.

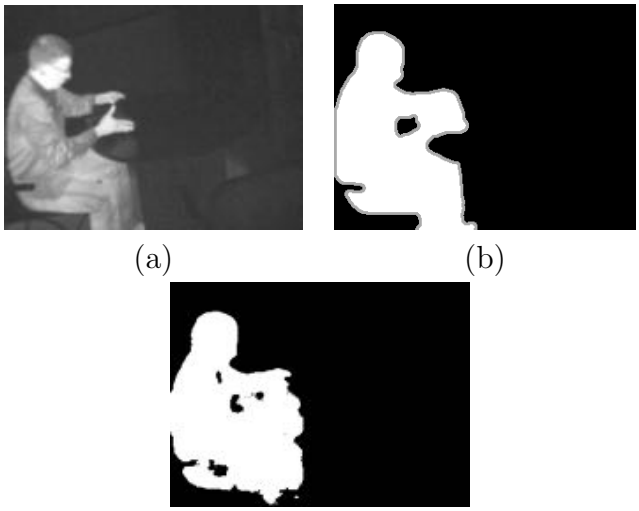
Chapter 5

Multi-Scale Deep Learning Architecture based Background Subtraction for Moving Object Detection

5.1 Introduction

The accuracy of the moving object detection using the background subtraction (BGS) techniques depends on the background construction modelling. However, background construction is a challenging task, as a video scene is generally, affected by illumination variation, shadow, disturbed weather, poor texture, low resolution, camera motion (jitter, tilting, and zooming), etc. Also, most of the existing BGS methods are scene-specific, and the outcomes of many algorithms are based upon manual parameter tuning. Further, the accuracy of these conventional techniques depends on hand-crafted features. Additionally, it is found that the visual surveillance-based BGS techniques are demonstrated for thermal videos. However, as thermal videos are mostly affected by the low resolution or missing information details, the SOTA techniques are found to be indigent in nature. Figure 5.1 (a) and (b) portray original image and corresponding ground-truth image considered for experimentation. The detected result obtained by an existing deep learning BGS scheme BSUV-Net 2.0 [143] is presented in Figure 5.1 (c). It may be observed from Figure 5.1 (c) that, the result obtained by the said technique is generate many isolated points in the

BGS and unable to provide the exact shape of the moving object.



(a) (b)

(c)

Figure 5.1: Visual analysis of (a) original image, (b) ground-truth image, and (c) moving object detection result obtained by the BSUV-Net 2.0 technique.

In this context, we have proposed two multi-scale deep learning architectures based background subtraction techniques for moving object detection in this chapter: modified ResNet-152 network with hybrid pyramidal pooling and multi-scale contrast preserving deep learning architecture. In the proposed modified ResNet-152 network with hybrid pyramidal pooling algorithm, a pre-trained modified ResNet-152 network is adhered to as an encoder with a transfer learning mechanism is capable of retaining deep features at various levels. Here, we designed a multi-scale features extraction (MFE) architecture which is a hybridization of pyramidal pooling architecture (PPA) and various atrous convolutional layers to extract multi-scale and multi-dimensional features at various scales. The decoder network consisting of stacked transposed convolution layers (Tconvs) can effectively projects the feature-level into the pixel-level. Again in the proposed, multi-scale contrast preserving deep learning architecture, the encoder network considers hybrid of convolution and atrous convolution blocks to preserve both sparse and dense features of a video with skip connections. The proposed encoder with the multi-scale contrast preservation block is able to keep multi-scale contrast features with less training loss. Here, the decoder network accurately projects the extracted features at different layers into pixel-level.

The proposed schemes are tested on benchmark databases: *changedetection.net*, and *Tripura University Video Dataset at Night Time (TU-VDN)*. The effectiveness of the pro-

posed modified ResNet-152 network with hybrid pyramidal pooling technique is validated against thirty-one state-of-the-art techniques. Further, the efficacy of the proposed multi-scale contrast preserving deep learning architecture is corroborated against twenty-eight existing SOTA techniques and is found to be effective. To confirm our findings, we have used qualitative and quantitative analysis.

The rest of the chapter is organized as follows. Section 5.2 describes the proposed deep learning based background subtraction schemes. The results and discussions with future works are carried out in Section 5.3. Section 5.4 draws the conclusions of the proposed works.

5.2 Proposed Multi-Scale Deep Learning Architecture based Background Subtraction for Moving Object Detection

In this chapter, we have proposed two multi-scale deep learning architectures for moving object detection: modified ResNet-152 network with hybrid pyramidal pooling and multi-scale contrast preserving deep learning architecture to detect the objects accurately with reduced missed/false alarms .

5.2.1 Proposed Modified ResNet-152 Network with Hybrid Pyramidal pooling

In this chapter, we have proposed a robust and stable encoder-decoder network for detecting the moving objects from a video scene with different challenging scenarios. Here, we proposed a deep CNN architecture as an encoder integrated with the multi-scale features extraction (MFE) block. We proposed a stacked transposed convolutional network as a decoder network to attain the said objective. The block diagram of the proposed scheme is presented in Figure 5.2.

5.2.1.1 The Encoder Configuration

In this work, a pertained ResNet-152 network is adhered to as the encoder. ResNet-152 [127] network operates on a deep residual learning framework and is widely used for several

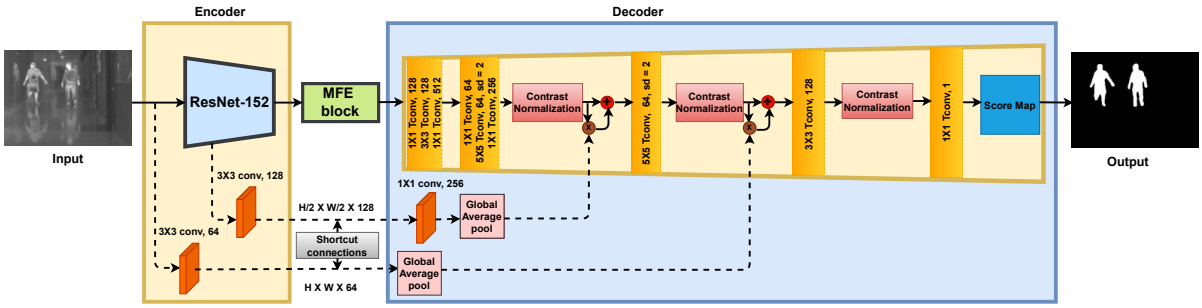
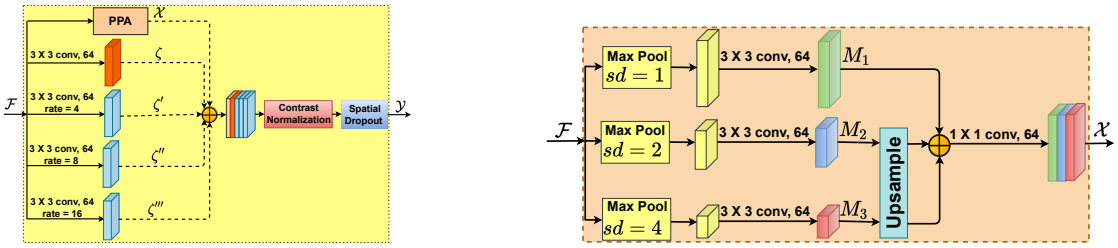


Figure 5.2: Block diagram of the proposed modified ResNet-152 network with hybrid pyramidal pooling scheme.



(a) (b)

Figure 5.3: Block diagram: (a) proposed multi-scale features extraction block and (b) proposed pyramidal pooling architecture.

Computer Vision applications. However, the said architecture was never explored to date for local change detection. In this article, we made an attempt to exploit the capabilities of the ResNet-152 network for local change detection. The ResNet-152 network consists of five blocks, where each block has a stacked of convolutional layers, batch normalization layers, and a rectified linear unit (ReLU) as an activation function. The Convolutional layers of the ResNet-152 network are capable of extracting the spatial information of the input image by using convolutional kernels. The batch normalization [144] layers are utilized in the deep neural network, which boosts the training speed with a faster learning rate. The use of ReLU in the network makes it faster and efficient.

In this work, we tried with a different variant of CNN architecture and found that ResNet-152 network to be a stable and efficient one. The motivation behind choosing the ResNet-152 network in the proposed scheme against another variant of CNN architecture is as follow:

(1) It may be observed that the deeper blocks of the ResNet-152 network gradually learn more complex features that provide improved performance.

(2) The lower blocks of the ResNet-152 network can learn and extract high spatial

resolution features with low-level local features: edges, colors, and textures, while the deeper blocks learn and extract high-level global features like objects and events with lower spatial resolution.

(3) The computational complexity of the ResNet-152 network is quite less with a higher number of layers.

(4) The ResNet-152 network uses the stacked residual networks with identity mapping to tackle the vanishing gradient problem and hence, provides better accuracy.

Here we have adhered to a modified form of deep Resnet-152 network in the proposed algorithm, which incorporates the initial three blocks. We keep the first two blocks weights same as the pre-trained ResNet-152 network, and the weights for the third block are learned using transfer learning. Transfer learning is a mechanism that converges knowledge from the source domain to the target domain. The use of transfer learning in the proposed scheme explores new tasks that depend on formerly acquired jobs by the deep ResNet-152 network. Additionally, it makes the model accurate, faster while training on fewer samples. We remove the max-pooling layer between the first two blocks to increase the use of deep multi-layer features with high spatial resolution. Also, we remove the fourth and fifth blocks of the ResNet-152 network to enhance the use of high spatial resolution and high-frequency components in the proposed scheme. The low-level features are extracted at the first block of the encoder by using 3×3 convolution layers with 64 and 128 filters. These low-level features are propagated towards the decoder network through shortcut connections followed by global average pooling that improves the feature representation.

5.2.1.2 The Multi-Scale Features Extraction (MFE) Block

In this work, We have proposed an MFE block to be sandwiched between the encoder-decoder network to preserve the contextual information from the deep multi-layer features. The deep feature maps ($\mathcal{F}$) of size $\frac{H}{4} \times \frac{W}{4} \times 512$ from the encoder network are given to an MFE block. Here $H \times W$ denotes the dimension of the input image. The MFE block is composed of a pyramidal pooling architecture (PPA) [145], and various atrous convolutional layers [146] as shown in Figure 5.3 (a). Additionally, the MFE block has a convolutional layer to preserve the sparse information. The contextual relationship among

the pixels of video scenes are the important elements in local change detection, and the lack of these may lead to many isolated points as missed/false alarms in the foreground map. Therefore, in this article, we have investigated a PPA that preserves the contextual relationship for video scenes. The graphical representation of the pyramidal pooling block is shown in Figure 5.3 (b), consisting of three max-pooling layers with strides (sd) of 1, 2 and 4; followed by 3×3 convolutional layers, with 64 filters. The max-pooling layer is utilized to preserve useful information in every pooling block and decrease the dimension of feature maps. The filter size for the max-pooling layers is considered to be 2×2 with strides of 1, 2, and 4, respectively. The output of the max-pooling layers followed by convolution layers are represented as M_1, M_2, and M_3. The dimension of M_1 same as input deep feature maps $\mathcal{F}$, whereas the dimensions of M_2 and M_3 are reduced by a factor of 2 and 4, respectively. Therefore, the dimensions of M_2 and M_3 are up-sampled and concatenated with M_1 along the depth dimension. The output of the PPA ($\mathcal{X}$) followed by activation function (ReLU) and 1×1 convolutional layer with 64 filters are given as,

$$\mathcal{X} = K_n * (M_1 \oplus (\uparrow \{M_2\}) \oplus (\uparrow \{M_3\})). \tag{5.1}$$

where K_n indicates the n number of kernels, $n \in \{1, 2 \cdots, 63, 64\}$. '$*$' represents the convolutional operation. '$\oplus$' is the concatenation operation, and '$\uparrow$' is the upsampling operation.

Again, feature maps $\mathcal{F}$ from the encoder are given to a 3×3 convolutional layer with 64 filters which will try to preserve the sparse information of the deep features and three 3×3 convolutional layers with different dilation rates. Here the convolutional layers with different dilation rates are called as atrous convolution, which retains the dense information of the deep features. The feature maps from the convolutional layer followed by ReLU are represented as ζ. For the target scenes, it is essential to use the spatial information of the current pixel with its neighborhood pixels to extract the dense contextual features. Therefore, atrous convolution is a mechanism that allows us to expand the receptive field of filters without increasing the parameters as well as computational complexity and can be defined as,

$$K_e = K + (K - 1)(R_d - 1), \tag{5.2}$$

where $K \times K$ indicates the filter size, R_d is the dilation rate, and $K_e \times K_e$ is the size of the expanded receptive field. Therefore, in the proposed MFE block, we precisely utilize three branches of 3×3 convolution layers each consisting of 64 filters, with dilation rates of 4, 8, and 16. The output of the atrous convolution layers followed by ReLU are represented as ζ', ζ'', and ζ'''.

Feature maps from the PPA, a convolutional layer, and three atrous convolutional layers are concatenated along the channels to generate 5×64 depth feature maps. The output multi-scale feature maps of an MFE block are obtained by processing these feature maps through the contrast normalization (CN) followed by an activation function (ReLU) and a spatial dropout (SD) layer. It may be observed that the use of contrast normalization instead of batch normalization enhances the performance of an MFE block. Also, the use of ReLU followed by SD with a rate of 0.25 increases the learning performance of the proposed model with fewer training samples. The output multi-scale feature maps from an MFE block ($\mathcal{Y}$) can be defined as

$$\mathcal{Y} = SD\{CN\{\mathcal{X} \oplus \zeta \oplus \zeta' \oplus \zeta'' \oplus \zeta'''\}\}. \tag{5.3}$$

where $SD\{\cdot\}$ indicates the spatial dropout operation and $CN\{\cdot\}$ is the contrast normalization operation.

5.2.1.3 The Decoder Configuration

The proposed decoder network efficiently decodes the MFE block's output that generates a dense probabilistic mask. Multi-scale feature maps from an MFE block ($\mathcal{Y}$) of size $\frac{H}{4} \times \frac{W}{4} \times 320$ are given to the decoder network for decoding, which consists of five blocks namely: TC1, TC2, TC3, TC4, and TC5. The TC1 block contains stack of two 1×1 and a 3×3 transposed convolutional layers with stride of 1. The main motive of the 1×1 transposed convolutional layer is to project the higher dimensional feature space into the lower dimensions. Also, the use of 1×1 transposed convolutional layer reduces the computational complexity of the network. Further, the use of the 3×3 transposed convolutional layer can extract the spatial information of the feature maps. Therefore, the stacked 1×1 and 3×3 transposed convolutional layers produce a robust network while performing on fewer parameters.

Initially, the features from an MFE block operated with TC1 block to project the

feature space from $\frac{H}{4} \times \frac{W}{4} \times 320$ to $\frac{H}{4} \times \frac{W}{4} \times 512$. TC2 block consists of a similar arrangement as TC1 block of layers except that we utilize 5×5 instead of 3×3 transposed convolutional layer with stride $sd = 2$. This block projects the feature space from $\frac{H}{4} \times \frac{W}{4} \times 512$ to $\frac{H}{2} \times \frac{W}{2} \times 256$. The feature maps from the TC2 block followed by contrast normalization (CN) and ReLU function are fused with low-level features to boost the features representation. These low-level features are pulled at the end of the first block of the encoder by utilizing a 3×3 convolutional layer with 128 filters. Subsequently, these features are propagated towards the decoder through shortcut connection followed by 1×1 convolutional layer, 256 filters with global average pooling (GAP). The use of GAP improves the performance of the model, which is robust to spatial translations of the low-level features. Then, the fused feature maps are given to the TC3 block, which consists of a 5×5 transposed convolutional layer, 64 filters with stride of 2. TC3 block is used to expand the feature maps to a size of $H \times W \times 64$.

The feature maps from the TC3 block followed by CN and ReLU function are fused with low-level features that are pulled from the beginning of the first block of an encoder. These low-level features are generated using a 3×3 convolutional layer with 64 filters, guided towards the decoder through shortcut connection followed by GAP. Further, the fused features are processed through the TC4 block, followed by CN and ReLU function, which project 64 feature maps to 128 feature maps. We have observed that these 128 feature maps in this block provide better representation to each pixel and improve the efficacy of the proposed model. Finally, in the TC5 block, we project 128 feature maps to 1 feature map by utilizing 1×1 transposed convolution layer with stride of 1 followed by a sigmoid activation function. The TC5 block generates a score map that predicts each pixel value's probability in the range [0, 1]. Later, we apply a threshold value to classify each pixel on the score map, either belongs to foreground or background.

Note that we add ReLU non-linearity to every transposed convolutional layer in TC1, TC2, and TC3. Also, to mitigate over-fitting in the proposed model, we use L2 regularization to the weights of the first layer of TC1, TC2, and 5×5 transposed convolutional layer of TC3 block.

5.2.1.4 Training Details and Parameter Settings

The proposed model is implemented in Keras framework with Tensorflow backend. Training is performed end-to-end over NVIDIA Tesla T4 GPU system with batch size of 2. The smaller batch size in the proposed scheme can converge the model faster and remarkable regularization effect. The model is trained using $N = 200$ frames with P number of pixels in each frame which can be given as,

$$\{\{L_a^b, C_a^b\}_{a=1}^P\}_{b=1}^N, a \in \{1, 2, \cdots P\}, b \in \{1, 2, \cdots N\}, \tag{5.4}$$

where L_a^b denotes the a^{th} pixel in b^{th} frame. The term C_a^b represents actual class of the a^{th} pixel of b^{th} frame.

Further, to compare the actual and predicted class label of each pixel, we use the binary cross-entropy loss ($BCEL$) function to train the model as;

$$BCEL = -\frac{1}{P}\sum_{a=1}^{P}[z_a log(\hat{z}_a) + (1 - z_a) log(1 - \hat{z}_a)], \tag{5.5}$$

where P is the number of training image pixels, $z_a \in \{0, 1\}$ is the actual label of pixel a, and $\hat{z}_a$ is the predicted pixel label. The function log indicates the logarithmic operation.

We have utilized *RMSProp* optimizer with $\rho = 0.9$ and $\epsilon = 1e - 08$ for training the proposed model. This provides a faster convergence rate as compared to other classical optimizers. Initially, the learning rate is fixed to 0.0001. If the validation loss during 5 consecutive epochs does not improve, the learning rate is further scaled down by 10. We kept a maximum of 100 epochs to train the model but the early stopping mechanism is adopted; if the validation loss does not improve for 10 successive epochs. If the training frames are fed to the model sequentially, it may result in biased learning weights. This problem arises because consecutive frames are highly correlated with each other. Therefore, to train the model initially, we shuffle the training frames. Further, these frames are divided into 90% as training and 10% for validation. We provide more weights to the foreground class and fewer weights to the background class to reduce the imbalanced data classification problem during training the model. We fix the $L2$ regularization strength to 0.0005 for the proposed model that can reduce the chance of over-fitting.

5.2.2 Proposed Multi-Scale Contrast Preserving Deep Learning Architecture

In this chapter, we have developed a BGS technique consisting of an efficient encoder-decoder deep-learning architecture. Local change detection is a challenging task because of the high uncertainty in the video scenes. As the deep learning network can extract the features in-depth and handle the vagueness in the challenging scenes, we have developed an encoder network using skip connection that provides better accuracy with increasing depth. The proposed multi-scale contrast preservation block can retain discriminative details from the in-depth features and act as a better feature representation block. The decoder network in the proposed model precisely projects the feature space into image space. The graphical representation of the proposed CNN architecture is presented in Figure 5.4 (a). The steps of the proposed scheme are narrated as follows. The proposed scheme contains three blocks: the encoder, the multi-scale contrast preservation, and the decoder.

5.2.2.1 The Encoder Configuration

In this work, we have proposed a novel encoder network with residual connections consisting of four blocks as shown in Figure 5.4 (a). Each block of the encoder network comprises five modules: module 1 to module 5 as shown in Figure 5.4 (b). The main motivation behind the proposed encoder is to model the sparse and dense features from the video frames using CNN models. We adhered to use one convolutional and four atrous convolutional [146] layers in each block of the proposed encoder model. The convolutional layer is responsible to extract the sparse features whereas the atrous convolution layers are responsible for extracting dense features. The atrous convolution are considered to be with dilation rates of 2, 4, 8, and 16 for extracting more dense features followed by contrast normalization (CN) layer [147] and activation function (ReLU). The module 1 for Block 1 consists of a stack of two 3×3 convolutional layers with 64 filters.

The use of CN layer in the proposed model achieves an improved performance against batch normalization for smaller batch size [147]. Further, the presence of the ReLU function makes the model faster. In addition to this, module 1 for Block 2 consists of a stack of two 3×3 convolutional layers with 128 filters and a 1×1 convolutional layer. The stacked 3×3 and 1×1 convolutional layers produce a robust network while

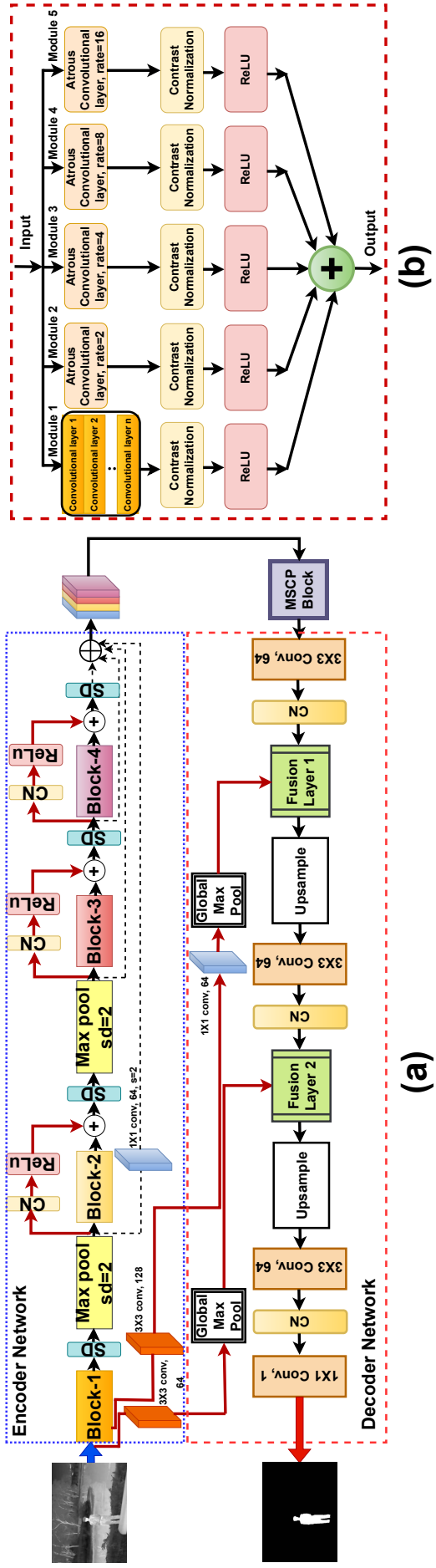


Figure 5.4: Block diagram: (a) proposed multi-scale contrast preserving deep learning architecture and (b) sample block of the proposed encoder.

performing fewer parameters. Further, module 1 for Block 3 consists of a stack of three 3×3 convolutional layers with 256 filters and a 1×1 convolutional layer with 64 filters followed by a CN layer and ReLU activation function. Finally, module 1 of Block 4 consists of a stack of three 3×3 convolutional layers with 512 filters and a 1×1 convolutional layer with 64 filters followed by a CN layer and ReLU.

For the same input, the distinct outcomes of all the modules are added and driven from block-1 to block-2 through the spatial dropout (SD) layer [148] with a rate of 0.25 and a max-pooling layer. Again, the same is propagated from block-1 to block-2 through skip connection followed by CN and ReLU layer and combined with the output of block-2. Similarly, the data transferred is accomplished from block-2 to block-3 but block-3 to block-4 is done without a max-pooling layer. The skip connection in the proposed model transfers the fine details of the image from the lower block to the higher block, producing coarse details of the image. A combination of fine and coarse details can promote the feature representation and will enhance the accuracy of the proposed model. We have used the max-pooling layer after the SD layer of block-1 and 2 to reduce the spatial dimension of input features that decrease the computational complexity of the model. The filter size for the max-pooling layer is 2×2 with stride of 2. Finally, the outputs of the block-1 followed by 1×1 convolutional layer with 64 filters, stride of 2, block-2, block-3 and block-4 are concatenated along the channels that represent the feature maps from the encoder ($\mathcal{F}$) and can be given as

$$\mathcal{F} = EB1 \oplus EB2 \oplus EB3 \oplus EB4. \tag{5.6}$$

where $EB1$, $EB2$, $EB3$, and $EB4$ denotes the output of four different encoder blocks. '$\oplus$' indicates the concatenation operation.

5.2.2.2 The Multi-Scale Contrast Preservation Block (MSCPB)

We have proposed a new multi-scale contrast preservation (MSCP) block as shown in Figure 5.5 to preserve the multi-scale sparse and dense features extracted in the proposed encoder. The MSCP block consists of a max-pooling layer followed by 1×1 convolutional layer with 64 filters, a 3×3 convolutional layer, and various atruos convolutional layers with dilation rate 4, 8, 16 with each consisting of 64 filters followed by kernel size 3×3 avg-pooling layer. For the encoder's deep feature, the max-pooling layer is used to retain

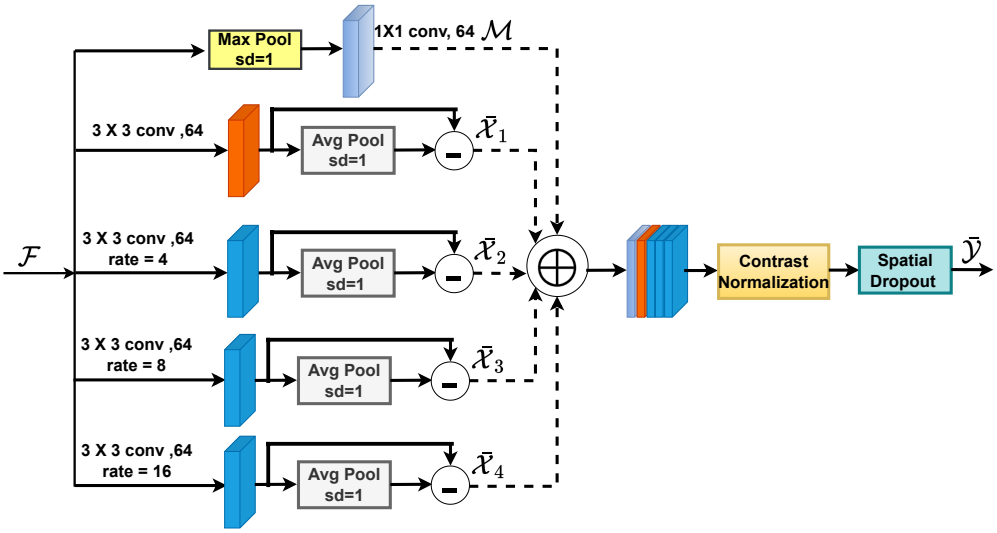


Figure 5.5: Block diagram of the proposed multi-scale contrast preservation block

the prominent detail in every pooling area and the output of this layer followed by 1×1 convolutional layer and ReLU function represented by $\mathcal{M}$. Since the local change detection is considered as a binary classification task, the outcome shows enormous contrast between the moving objects and the background. Hence, in this work, we have preserved the contrast details $\bar{\mathcal{X}}_u$ of the encoder's feature and can be given as;

$$\bar{\mathcal{X}}_u = ReLU\{CV_u - AvgPool(CV_u)\}. \tag{5.7}$$

where CV_u denotes the feature maps of a convolutional layer and various atrous convolutional layers, $u \in \{1, 2, 3, 4\}$.

The output of the MSCP block ($\bar{\mathcal{Y}}$) is generated by concatenating $\mathcal{M}$ and $\bar{\mathcal{X}}_u$ along the depth dimension and processed through the contrast normalization (CN) followed by an activation function (ReLU) and a spatial dropout (SD) layer with a dropout rate of 0.25. The multi-scale feature maps from the MSCP block can be calculated as;

$$\bar{\mathcal{Y}} = SD\{CN\{\mathcal{M} \oplus \bar{\mathcal{X}}_u\}\}. \tag{5.8}$$

5.2.2.3 The Decoder Configuration

The decoder network consists of three blocks where each block sandwiches a 3×3 convolutional layer with 64 filters, contrast normalization (CN) layer, and ReLU activation function. The multi-scale feature maps from the MSCP block are given to the decoder

network's first block, which produces 64 feature maps. The output of this block is fused with low-level features, which is pulled from the end of module 1 of the encoder first block using a 3×3 convolutional layer with 128 filters. These low-level features are propagated from the encoder to the decoder network through skip connection followed by 1×1 convolutional layer with 64 filters and global max pooling layer. The fusion layer 1 in the decoder enhances the feature representation where the low-level features are initially multiplied with the output of the decoder first block and subsequently combined with the initial feature maps obtained by the same. Finally, these fused feature maps are up-sampled and given to the next block. Here, a similar kind of operation is performed, and fusion layer 2 boosts the feature representation. Here, for fusion, the low-level features are extracted using a 3×3 convolutional layer with 64 filters from the beginning of the encoder first block and propagated towards the decoder followed by skip connection and global max pooling layer. The fused feature maps from fusion layer 2 are up-sampled and fed to the third block of a decoder, which projects the fused features into 64 feature maps. The output of this block is followed by a 1×1 convolutional layer with 1 filter and a sigmoid function that provides a single feature map. We have applied a threshold value of 0.9 on the single feature to predict, each pixel either belongs to foreground or background.

5.2.2.4 Training Details and Parameter Settings

The proposed model is implemented in Keras framework with Tensorflow backend. Training is performed end-to-end over NVIDIA Tesla T4 GPU system with batch size of 1. The smaller batch size in the proposed scheme can converge the model faster and remarkable regularization effect. The model is trained using $N = 200$ frames. We have utilized *RMSProp* optimizer with $\rho = 0.9$ and $\epsilon = 1e - 08$ for training the proposed model. This provides a faster convergence rate as compared to other classical optimizers. Initially, the learning rate is fixed to 0.0001. If the validation loss during 5 consecutive epochs does not improve, the learning rate is further scaled down by 10. We kept a maximum of 100 epochs to train the model but the early stopping mechanism is adopted; if the validation loss does not improve for 10 successive epochs. If the training frames are fed to the model sequentially, it may result in biased learning weights. This problem arises because consecutive frames are highly correlated with each other. Therefore, to train

the model initially, we shuffle the training frames. Further, these frames are divided into 90% as training and 10% for validation. We provide more weights to the foreground class and fewer weights to the background class to reduce the imbalanced data classification problem during training the model.

5.3 Results and Discussions

Both of the proposed techniques are executed on a Core $i7$ system with 16 GB RAM, 1.5 MB L2 cache. The proposed techniques are implemented by python programming with the Windows−10 operating system. The training and testing of the models are performed by utilizing $NVIDIA$ Tesla $T4$ GPU provided by Google Colaboratory with $Keras$. The proposed algorithms are validated by testing it with two benchmark databases: *TU-VDN* [78], and *changedetection.net* [113]. To validate the effectiveness of the proposed techniques, we have carried out both qualitative and quantitative analysis on all the considered sequences. The performance of the proposed techniques are verified by comparing the results obtained by it with several state-of-the-art BGS techniques. The performance of the proposed techniques are corroborated through various quantitative assessments: average Precision, average Recall, average F-measure, average percentage of wrong classifications [114], average Matthews correlation co-efficient [78], and average accuracy [78].

5.3.1 Qualitative illustration of Modified ResNet-152 Network with Hybrid Pyramidal pooling

The visual analysis of the change detection results are initially, carried out using few testing sequences taken from the *changedetection.net* dataset: PeopleInShade, DiningRoom, BusStation, Highway, Sofa and LakeSide. These testing sequences incorporate diverse challenging effects: hard and soft shadow, ghost artifacts, noise, low contrast, high uncertainty and high ambiguity pixel values, etc. For visual illustration of the proposed technique is visually compared against few existing state-of-the-art deep learning based BGS techniques: DeepBS [101], BSPVGAN [112], WisenetMD [140], Cascade CNN [103], IUTIS-5 [141], BSUV_Net [108], FgSegNet_S_FPM [149] and FgsegNet_v2 [109].

All the original frames and the corresponding ground-truth images of the considered six challenging sequences are shown in Figure 5.6 (a) and (b). The results obtained by

the DeepBS scheme as shown in Figure 5.6 (c), illustrate that many of the edge pixels are missing due to imbalanced pixel values in various video frames. Thus, the DeepBS technique provides many missed alarms in the change detection results. Figure 5.6 (d) shows the segmented foreground results obtained by the BSPVGAN technique where many false alarms appear in the scene. The object detection results achieved by the WisenetMD method are given in Figure 5.6 (e), where it can be perceived that few details of the moving objects are missing. Hence, the missed alarm rate is high in the detected results. Figure 5.6 (f) represents the Cascade CNN results where many false detections are observed. Figure 5.6 (g) shows the results of the IUTIS-5 scheme, which is unable to identify the small variation in grey values and generate ghosts in the challenging scene. The results obtained by the BSUV_Net technique are presented in Figure 5.6 (h). In the said method, it can be seen that many parts of the background is identified as foreground. The results obtained by FgSegNet_S_FPM and FgsegNet_v2 are shown in Figure 5.6 (i) and (j), where some better performance is obtained for all the considered six sequences. However, the said techniques wrongly classified some object pixels as background. On the contrary, the results obtained by the proposed algorithm presented in Figure 5.6 (k) yields better performance against other state-of-the-art techniques as the foreground and background pixels are accurately classified. The proposed technique precisely detects the shape of the moving objects as compared to other considered existing state-of-the-art techniques.

The visual analysis of the moving object detection are also carried out using various testing sequences taken from *TU-VDN* dataset: Rain, Lowlight, Fog and Dust with key challenges like flat cluttered background and dynamic background. The considered original frames and the corresponding ground-truth images are presented in Figure 5.7 (a)-(b). The results on five competitive techniques: SuBSENSE [72], LOBSTER [150], PBAS [151], KDE [56], and VuMeter [152] are shown in Figures 5.7 (c)-(g), respectively. It may be observed that the results obtained by the SOTA techniques are unable to detect the objects accurately. Also, these SOTA techniques generates holes and false alarms in the detected results. However, the results obtained by the proposed algorithm is shown in Figure 5.7 (h) confirm our findings by giving better results than the SOTA techniques.

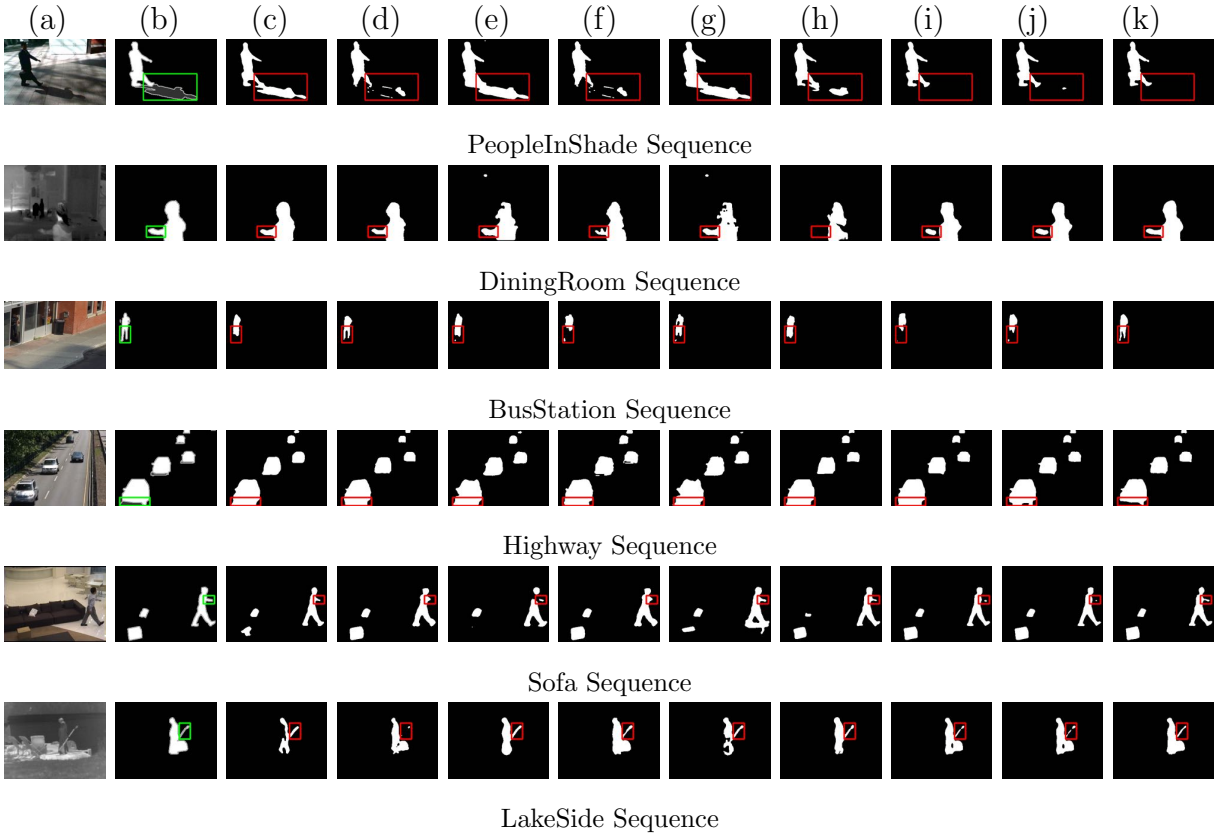


Figure 5.6: Moving object detection for different sequences: (a) original frame (b) corresponding ground-truth, moving object detection results obtained by deep learning based BGS schemes: (c) DeepBS, (d) BSPVGAN, (e) WisenetMD, (f) Cascade CNN, (g) IUTIS_5, (h) BSUV_Net (i) FgSegNet_S_ FPM, (j) FgSegNet_v2 and (k) proposed modified ResNet-152 network with hybrid pyramidal pooling scheme.

5.3.2 Quantitative comparison of Modified ResNet-152 Network with Hybrid Pyramidal pooling

The performance of the proposed technique is corroborated through various quantitative assessment: average Precision, average Recall, average F-measure, average PWC [114], average Matthews correlation co-efficient, and average accuracy [78]. The effectiveness of the proposed algorithm is validated by testing it with two benchmark databases: *TU-VDN* [78], and *changedetection.net* [113].

At the first stage of the experiment, we have used *TU-VDN* [78] to check the efficiency of the proposed technique. This database consists of various sequences that are captured under various atmospheric conditions such as dusty, rainy, and foggy with key challenges like flat cluttered background and dynamic background. For this database, the efficacy of proposed scheme is evaluated against fifteen state-of-the-art techniques: ViBe [70], SuBSENSE [72], LOBSTER [150], PAWCS [136], FST [153], PBAS [151], Multicue [154],

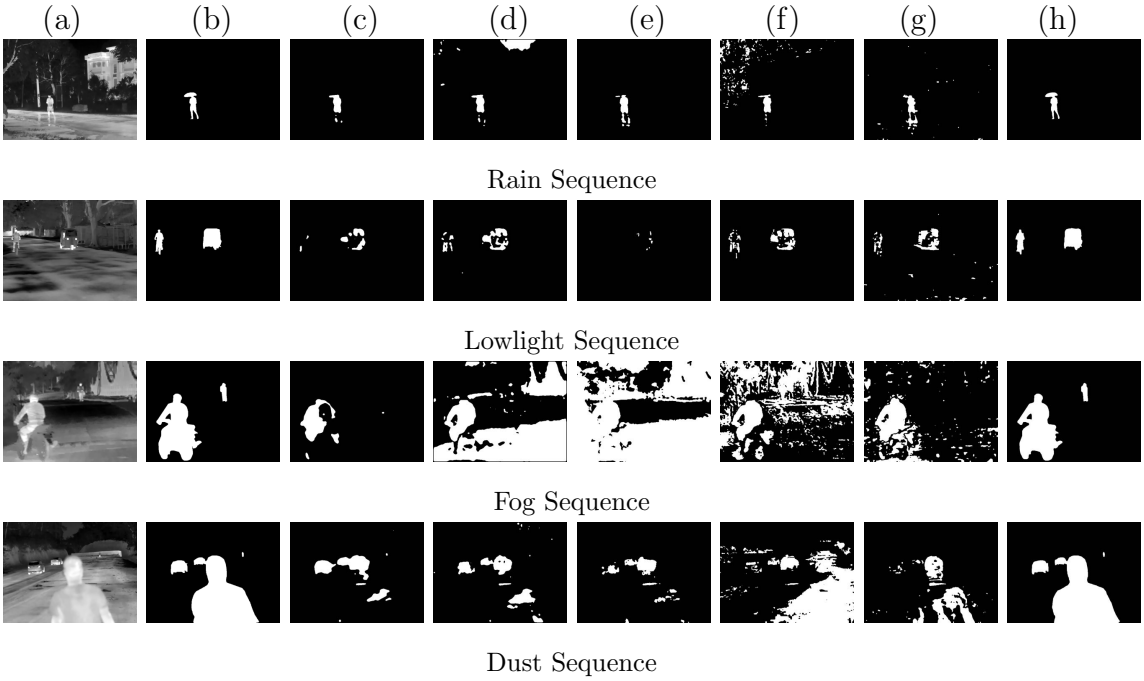


Figure 5.7: Moving object detection for different sequences: (a) original frame (b) corresponding ground-truth, moving object detection results obtained by non-deep learning based BGS schemes: (c) SuBSENSE, (d) LOBSTER, (e) PBAS, (f) KDE, (g) Vumeter and (f) proposed modified ResNet-152 network with hybrid pyramidal pooling scheme.

ISBM [155], MTD [156], VuMeter [152], KDE [56], MoG_V2 [157], Eigenbackground [158], Codebook [159], and ALWBP [78]. The evaluation of the proposed scheme is carried out using three quantitative measures: average F-measure, average Matthews correlation coefficient, and average accuracy [78] are provided in Table 5.1. From Table 5.1, it may be found that except accuracy measure for fog sequence; the proposed scheme provides a higher accuracy by all these measures on different categories as compared to other considered SOTA techniques.

In the next stage of the experiment, we have used *changedetection.net* database, which consists of eleven categories of video with fifty-three various challenging sequences. Here, we have considered a Recall, Specificity, False-positive rate (FPR), False-negative rate (FNR), Precision, F-measure and PWC are the quantitative evaluation metrics. The above-mentioned evaluation measures for results obtained by the proposed scheme on *changedetection.net* are reported in Table 5.2. It can be observed from Table 5.2 that the proposed technique provides adequate values of all the above-mentioned measures.

To verify the efficiency of the proposed scheme, we compared the results obtained by it on five thermal sequences available at *changedetection.net* database against eight deep learning-based BGS techniques: DeepBS [101], BSPVGAN [112], WisenetMD [140], Cas-

cade CNN [103], IUTIS-5 [141], BSUV_Net [108], SemanticBGS [142], and BSUV_Net2.0 [143], and nine non-deep learning-based state-of-the-art BGS techniques: KDE [56], GMM [55], PAWCS [136], SuBSENSE [72], SOBS-CF [91], WeSamBE [75], Multimode Background [74], Shared Model [137] and Spectral-360[138]. From Table 5.3, it may be observed that the proposed scheme has provided a higher accuracy in terms of all the considered measures than the SOTA techniques. Similarly, in order to justify the effectiveness of proposed scheme, we compared the results obtained by it on all the available videos (53 videos) available at *changedetection.net* database against nineteen state-of-the-art-techniques: Ten deep learning and nine deterministic existing techniques. The ten deep learning based BGS techniques: DeepBS [101], BSPVGAN [112], WisenetMD [140], Cascade CNN [103], IUTIS-5 [141], BSUV_Net [108], SemanticBGS [142], FgSegNet_S_FPM [149], FgsegNet_v2 [109] and BSUV_Net2.0 [143]. Table 5.4 presents the evaluation measures on *changedetection.net* database obtained by the proposed technique and considered state-of-the-art deep learning based BGS techniques. It may observed from Table 5.4 that, the proposed background subtraction technique yields higher values of average Precision, average F-measure and lower value of average PWC as compared to all deep learning based state-of-the-art techniques. However, the proposed technique provides a closer average Recall value to the FgSegNet_S_FPM and the FgsegNet_v2 techniques.

The performance of the proposed technique further evaluated by comparing it against the considered non deep learning based architectures: KDE [56], GMM [55], PAWCS [136], SuBSENSE [72], SOBS-CF [91], WeSamBE [75], Multimode Background [74], Shared Model [137] and Spectral-360[138]. From Table 5.4, it can be observed that the proposed technique provides higher values of average Precision, average Recall, average F-measure, and lower value of average PWC for all the considered image sequences than the considered non deep learning based state-of-the-art techniques.

5.3.3 Qualitative illustration of Multi-Scale Contrast Preserving Deep Learning Architecture

The visual analysis of the moving object detection is carried out using various testing sequences taken from *changedetection.net* and *TU-VDN* are shown in Figures 5.8-5.9. The considered original frames and the corresponding ground-truth images from the *changedetection.net* are presented in Figure 5.8 (a)-(b). The DeepBS [101] based BGS scheme's

Table 5.1: Quantitative comparisons on all the sequences of TU-VDN database

Approaches	Dust			Fog			Rain			Low Light		
	Avg. F-Measure	Avg. MCC	Avg. ACC	Avg. F-Measure	Avg. MCC	Avg. ACC	Avg. F-Measure	Avg. MCC	Avg. ACC	Avg. F-Measure	Avg. MCC	Avg. ACC
ViBe [70]	0.5565	0.5823	0.9740	0.6844	0.7119	0.9921	0.7307	0.7379	0.9877	0.5738	0.5954	0.9907
SuBSENSE [72]	0.5405	0.5679	0.9722	0.8204	0.8218	**0.9969**	0.6112	0.6171	0.9788	0.4812	0.5091	0.9837
LOBSTER [150]	0.4967	0.5346	0.9688	0.6649	0.6769	0.9943	0.5658	0.5949	0.9793	0.5244	0.5413	0.9890
PAWCS [136]	0.2322	0.2872	0.9656	0.5176	0.5559	0.9945	0.6628	0.6752	0.9875	0.3155	0.3505	0.9870
FST [153]	0.2360	0.2564	0.7189	0.5173	0.5659	0.9677	0.6760	0.6938	0.9892	0.4061	0.4509	0.9523
PBAS [151]	0.4033	0.4311	0.9547	0.6936	0.7086	0.9952	0.7035	0.6893	0.9823	0.5668	0.5803	0.9901
Multicue [154]	0.6166	0.6345	0.9714	0.5511	0.5913	0.9717	0.5511	0.5912	0.9717	0.4961	0.5314	0.9798
ISBM [155]	0.2191	0.2443	0.7040	0.4951	0.5456	0.9501	0.2773	0.3387	0.9312	0.3594	0.4064	0.9654
MTD [156]	0.4709	0.4772	0.9729	0.5561	0.5640	0.9928	0.4843	0.4948	0.9802	0.4656	0.4927	0.9814
VuMeter [152]	0.2171	0.2578	0.9667	0.3071	0.3698	0.6083	0.6225	0.6373	0.9814	0.3336	0.3882	0.9862
KDE [56]	0.2689	0.2978	0.9453	0.5332	0.5522	0.9938	0.6198	0.6315	0.9606	0.3845	0.3996	0.9691
MoG_V2 [157]	0.2208	0.2799	0.9673	0.2117	0.2509	0.9928	0.3532	0.3963	0.9651	0.3119	0.3486	0.9854
Eigenbackground [158]	0.3603	0.3996	0.9184	0.3413	0.3734	0.9629	0.2120	0.1761	0.6499	0.3695	0.4190	0.9673
Codebook [159]	0.2066	0.2245	0.6807	0.2319	0.3299	0.9111	0.2736	0.3958	0.9021	0.3093	0.3623	0.8848
ALWBP [78]	0.7002	0.6983	0.9763	0.7451	0.7456	0.9960	0.6962	0.6832	0.9877	0.6555	0.6658	0.9891
Proposed	**0.9131**	**0.9093**	**0.9926**	**0.8706**	**0.8745**	0.9937	**0.8792**	**0.8817**	**0.9959**	**0.8681**	**0.8709**	**0.9919**

Table 5.2: Quantitative analysis on all the sequences of changedetection.net database

Category	Recall	Specificity	FPR	FNR	Precision	F-Measure	PWC
BadWeather	0.9832	0.9999	0.0001	0.0168	0.9921	0.9876	0.0354
Baseline	0.9968	0.9999	0.0001	0.0032	0.9982	0.9975	0.0147
Camera Jitter	0.9953	0.9999	0.0001	0.0047	0.9968	0.9960	0.0339
Dynamic Background	0.9949	0.9999	0.0000	0.0051	0.9957	0.9953	0.0065
Intermittent Object Motion	0.9915	0.9999	0.0001	0.0085	0.9963	0.9939	0.0616
Low Framerate	0.9433	0.9999	0.0001	0.0567	0.9613	0.9520	0.0262
Night Videos	0.9788	0.9997	0.0003	0.0212	0.9860	0.9824	0.0717
PTZ	0.9909	0.9999	0.0000	0.0091	0.9953	0.9931	0.0144
Shadow	0.9955	0.9999	0.0001	0.0045	0.9963	0.9959	0.0338
Thermal	0.9912	0.9998	0.0002	0.0088	0.9958	0.9935	0.0564
Turbulence	0.9768	0.9999	0.0001	0.0232	0.9813	0.9790	0.0222
Average	0.9853	0.9999	0.0001	0.0147	0.9905	0.9878	0.0343

Table 5.3: Quantitative comparisons on 5 thermal sequences of changedetection.net database

Approaches	Avg.Precision	Avg.Recall	Avg.F-Measure	Avg.PWC
DeepBS [101]	0.9257	0.6637	0.7583	3.5773
BSPVGAN [112]	0.9770	0.9763	0.9764	0.2406
WisenetMD [140]	0.8696	0.7867	0.8152	1.8993
Cascade CNN [103]	0.8577	0.9461	0.8958	1.0478
IUTIS-5 [141]	0.8969	0.7990	0.8303	1.1484
BSUV_Net [108]	0.8551	0.8739	0.8581	1.7058
SemanticBGS [142]	0.9118	0.7664	0.8219	1.3897
BSUV_Net2.0 [143]	0.9359	0.8594	0.8932	1.1659
KDE [56]	0.8974	0.6725	0.7423	1.6795
GMM [55]	0.8652	0.5691	0.6621	4.2642
PAWCS [136]	0.8280	0.8504	0.8324	1.4018
SuBSENSE [72]	0.8328	0.8161	0.8171	2.0125
SOBS-CF [91]	0.8715	0.6347	0.7140	1.8021
WeSamBE [75]	0.8554	0.7727	0.7962	2.3538
Multimode Background [74]	0.8268	0.8162	0.8194	1.4289
SharedModel [137]	0.8072	0.8618	0.8319	1.8656
Spectral-360 [138]	0.9114	0.7238	0.7764	1.6337
Proposed	**0.9958**	**0.9912**	**0.9935**	**0.0564**

results are presented in Figure 5.8 (c), which produces many false alarms. Figure 5.8 (d) displays the results obtained by the WisenetMD [140] based BGS scheme. It is observed from this figure that the WisenetMD based scheme produces ghosts for the considered frames. Figure 5.8 (e) represents the results of the Cascade CNN [103] based BGS scheme,

Table 5.4: Quantitative comparisons on all the sequences of changedetection.net database

Approaches	Avg.Precision	Avg.Recall	Avg.F-Measure	Avg.PWC
DeepBS [101]	0.8332	0.7545	0.7458	1.9920
BSPVGAN [112]	0.9472	0.9544	0.9501	0.2272
WisenetMD [140]	0.7668	0.8179	0.7535	1.6136
Cascade CNN [103]	0.8997	0.9506	0.9209	0.4052
IUTIS-5 [141]	0.8087	0.7849	0.7717	1.1986
BSUV_Net [108]	0.8113	0.8203	0.7868	1.1402
SemanticBGS [142]	0.8305	0.7890	0.7892	1.0722
FgSegNet_S_FPM [149]	0.9751	**0.9896**	0.9804	0.0461
FgsegNet_v2 [109]	0.9823	0.9891	0.9847	0.0402
BSUV_Net2.0 [143]	0.9011	0.8136	0.8387	0.7614
KDE [56]	0.5811	0.7375	0.5688	5.6262
GMM [55]	0.6025	0.6846	0.5707	3.7667
PAWCS [136]	0.7857	0.7718	0.7403	1.1992
SuBSENSE [72]	0.7509	0.8124	0.7408	1.6780
SOBS-CF [91]	0.5831	0.7805	0.5883	6.0709
WeSamBE [75]	0.7679	0.7955	0.7446	1.5105
Multimode Background [74]	0.7382	0.7389	0.7288	1.2614
SharedModel [137]	0.7503	0.8098	0.7474	1.4996
Spectral-360 [138]	0.7054	0.7345	0.6732	2.2722
Proposed	**0.9905**	0.9853	**0.9878**	**0.0343**

where many object pixels are misclassified as the background pixels. The results for the IUTIS_5 [141] based BGS scheme are reported in Figure 5.8 (f), where mainly false detections are observed. Figure 5.8 (g), shows the results of the BSUV_net [108] scheme; where a high false-positive rate is observed. Figure 5.8 (h) denotes the results obtained by the SemanticBGS [142], which is unable to detect the shape of the objects accurately. The results of BSUV_net2.0 [143] are presented in 5.8 (i), where poor results are observed. The results obtained by the proposed BGS scheme as shown in Figure 5.8 (j). It is found from this figure that the proposed algorithm is able to detect both object and background pixels with better accuracy.

The considered original frames and the corresponding ground-truth images from the *TU-VDN* are presented in Figure 5.9 (a)-(b). The results on five competitive techniques: SuBSENSE [72], LOBSTER [150], PBAS [151], KDE [56], and VuMeter [152] are shown in Figures 5.9 (c)-(g), respectively. It may be observed that the results obtained by the SOTA techniques have a high false-positive rate. Also, these SOTA techniques generates holes and missed alarms in the detected results. However, the results obtained by the proposed algorithm (shown in Figure 5.9 (h)) confirms our findings by giving better results against the SOTA techniques.

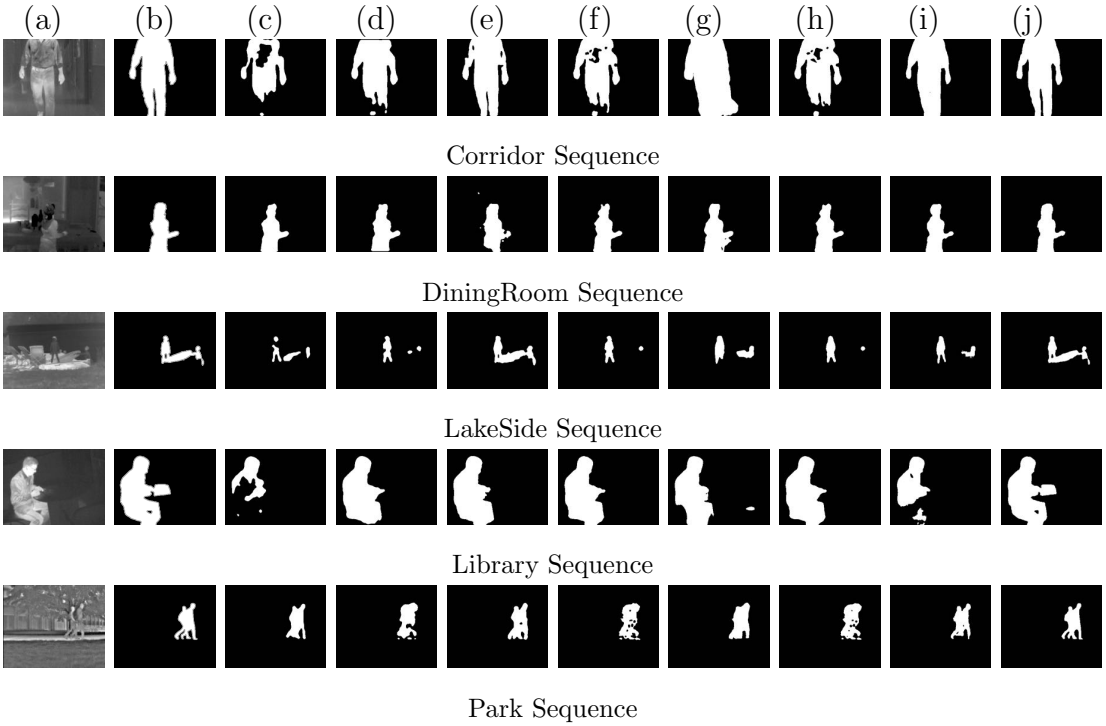


Figure 5.8: Moving object detection for different sequences: (a) original frame (b) corresponding ground-truth, moving object detection results obtained by deep learning based BGS schemes: (c) DeepBS, (d) WisenetMD, (e) Cascade CNN, (f) IUTIS_5, (g) BSUV_net (h) SemanticBGS, (i) BSUV_net2.0 and (j) proposed multi-scale contrast preserving deep learning architecture scheme.

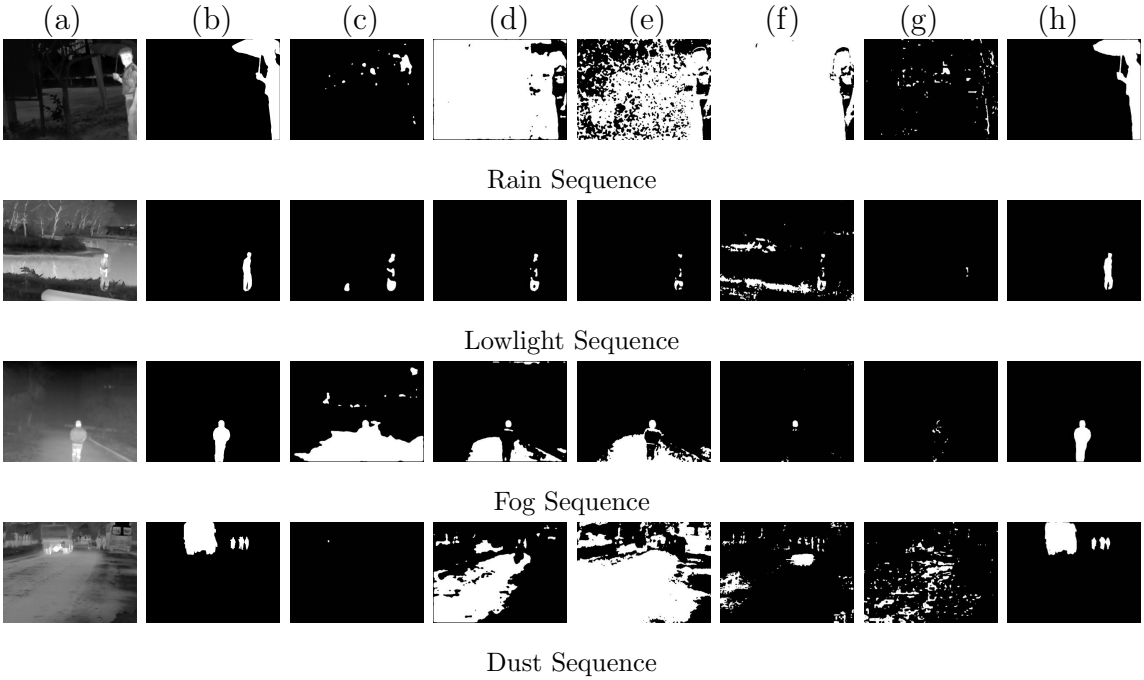


Figure 5.9: Moving object detection for different sequences: (a) original frame (b) corresponding ground-truth, moving object detection results obtained by non-deep learning based BGS schemes: (c) SuBSENSE, (d) LOBSTER, (e) PBAS, (f) KDE, (g) VuMeter and (h) proposed multi-scale contrast preserving deep learning architecture scheme.

5.3.4 Quantitative comparison of Multi-Scale Contrast Preserving Deep Learning Architecture

The evaluation of the proposed scheme is carried out using three quantitative measures: average F-measure, average Matthews correlation co-efficient, and average accuracy [78] for *TU-VDN* database. Similarly, for a fair evaluation on *changedetection.net* database, we have used four measures: average Precision, average Recall , average F-measure, and average PWC [99]. The evaluation of the performance of the proposed scheme on *TU-VDN* database is provided in Table 5.5. To evaluate the proposed scheme on *TU-VDN* database, we have used fifteen SOTA BGS techniques: ViBe [70], SuBSENSE [72], LOBSTER [150], PAWCS [136], FST [153], PBAS [151], Multicue [154], ISBM [155], MTD [156], VuMeter [152], KDE [56], MoG_V2 [157], Eigenbackground [158], Codebook [159], and ALWBP [78]. It may be observed that except accuracy measure for fog sequence; the proposed scheme provides a higher accuracy by all measures on different categories as compared to other considered SOTA techniques. The results of the proposed scheme on five thermal sequences available at *changedetection.net* database are provided in Table 5.6. The proposed scheme is compared against sixteen SOTA BGS techniques: DeepBS [101], WisenetMD [140], Cascade CNN [103], IUTIS-5 [141], BSUV_Net [108], SemanticBGS [142], BSUV_Net2.0 [143], KDE [56], GMM [55], PAWCS [136], SuBSENSE [72], SOBS-CF [91], WeSamBE [75], Multimode Background [74], Shared Model [137] and Spectral-360[138]. It may be observed that the proposed scheme has provided a higher accuracy in terms of all considered measures. We have also tested the proposed scheme on all the available videos (53 videos) available at *changedetection.net* and found that the proposed scheme performance is best as compared to the sixteen SOTA techniques (as provided in Table 5.7).

5.3.5 Discussions and Future Works

Background subtraction is an essential step in any surveillance system. Here, the ultimate goal is to detect the local changes, and the system could be employed to face many of the real-life challenges. However, foreground and background separation is a challenging task, as in general, a video scene is affected by illumination variation, shadow, disturbed weather, poor texture, poor resolution, camera motion (jitter, tilting, and zooming), etc.

Table 5.5: Quantitative comparisons on all the sequences of TU-VDN database

Approaches	Dust			Fog			Rain			Low Light		
	Avg. F-Measure	Avg. MCC	Avg. ACC	Avg. F-Measure	Avg. MCC	Avg. ACC	Avg. F-Measure	Avg. MCC	Avg. ACC	Avg. F-Measure	Avg. MCC	Avg. ACC
ViBe [70]	0.5565	0.5823	0.9740	0.6844	0.7119	0.9921	0.7307	0.7379	0.9877	0.5738	0.5954	0.9907
SuBSENSE [72]	0.5405	0.5679	0.9722	0.8204	0.8218	**0.9969**	0.6112	0.6171	0.9788	0.4812	0.5091	0.9837
LOBSTER [150]	0.4967	0.5346	0.9688	0.6649	0.6769	0.9943	0.5658	0.5949	0.9793	0.5244	0.5413	0.9890
PAWCS [136]	0.2322	0.2872	0.9656	0.5176	0.5559	0.9945	0.6628	0.6752	0.9875	0.3155	0.3505	0.9870
FST [153]	0.2360	0.2564	0.7189	0.5173	0.5659	0.9677	0.6760	0.6938	0.9892	0.4061	0.4509	0.9523
PBAS [151]	0.4033	0.4311	0.9547	0.6936	0.7086	0.9952	0.7035	0.6893	0.9823	0.5668	0.5803	0.9901
Multicue [154]	0.6166	0.6345	0.9714	0.5511	0.5913	0.9717	0.5511	0.5912	0.9717	0.4961	0.5314	0.9798
ISBM [155]	0.2191	0.2443	0.7040	0.4951	0.5456	0.9501	0.2773	0.3387	0.9312	0.3594	0.4064	0.9654
MTD [156]	0.4709	0.4772	0.9729	0.5561	0.5640	0.9928	0.4843	0.4948	0.9802	0.4656	0.4927	0.9814
VuMeter [152]	0.2171	0.2578	0.9667	0.3071	0.3698	0.6083	0.6225	0.6373	0.9814	0.3336	0.3882	0.9862
KDE [56]	0.2689	0.2978	0.9453	0.5332	0.5522	0.9938	0.6198	0.6315	0.9606	0.3845	0.3996	0.9691
MoG_V2 [157]	0.2208	0.2799	0.9673	0.2117	0.2509	0.9928	0.3532	0.3963	0.9651	0.3119	0.3486	0.9854
Eigenbackground [158]	0.3603	0.3996	0.9184	0.3413	0.3734	0.9629	0.2120	0.1761	0.6499	0.3695	0.4190	0.9673
Codebook [159]	0.2066	0.2245	0.6807	0.2319	0.3299	0.9111	0.2736	0.3958	0.9021	0.3093	0.3623	0.8848
ALWBP [78]	0.7002	0.6983	0.9763	0.7451	0.7456	0.9960	0.6962	0.6832	0.9877	0.6555	0.6658	0.9891
Proposed	**0.8349**	**0.8286**	**0.9864**	**0.8589**	**0.8631**	0.9929	**0.8057**	**0.8036**	**0.9934**	**0.8554**	**0.8548**	**0.9908**

Table 5.6: Quantitative comparison on 5 thermal sequences of changedetection.net database

Approaches	Avg.Precision	Avg.Recall	Avg.F-Measure	Avg.PWC
DeepBS [101]	0.9257	0.6637	0.7583	3.5773
WisenetMD [140]	0.8696	0.7867	0.8152	1.8993
Cascade CNN [103]	0.8577	0.9461	0.8958	1.0478
IUTIS-5 [141]	0.8969	0.7990	0.8303	1.1484
BSUV_Net [108]	0.8551	0.8739	0.8581	1.7058
SemanticBGS [142]	0.9118	0.7664	0.8219	1.3897
BSUV_Net2.0 [143]	0.9359	0.8594	0.8932	1.1659
KDE [56]	0.8974	0.6725	0.7423	1.6795
GMM [55]	0.8652	0.5691	0.6621	4.2642
PAWCS [136]	0.8280	0.8504	0.8324	1.4018
SuBSENSE [72]	0.8328	0.8161	0.8171	2.0125
SOBS-CF [91]	0.8715	0.6347	0.7140	1.8021
WeSamBE [75]	0.8554	0.7727	0.7962	2.3538
Multimode Background [74]	0.8268	0.8162	0.8194	1.4289
SharedModel [137]	0.8072	0.8618	0.8319	1.8656
Spectral-360 [138]	0.9114	0.7238	0.7764	1.6337
Proposed	**0.9718**	**0.9858**	**0.9787**	**0.5485**

Table 5.7: Quantitative comparisons on all the sequences of changedetection.net database

Approaches	Avg.Precision	Avg.Recall	Avg.F-Measure	Avg.PWC
DeepBS [101]	0.8332	0.7545	0.7458	1.9920
WisenetMD [140]	0.7668	0.8179	0.7535	1.6136
Cascade CNN [103]	0.8997	0.9506	0.9209	0.4052
IUTIS-5 [141]	0.8087	0.7849	0.7717	1.1986
BSUV_Net [108]	0.8113	0.8203	0.7868	1.1402
SemanticBGS [142]	0.8305	0.7890	0.7892	1.0722
BSUV_Net2.0 [143]	0.9011	0.8136	0.8387	0.7614
KDE [56]	0.5811	0.7375	0.5688	5.6262
GMM [55]	0.6025	0.6846	0.5707	3.7667
PAWCS [136]	0.7857	0.7718	0.7403	1.1992
SuBSENSE [72]	0.7509	0.8124	0.7408	1.6780
SOBS-CF [91]	0.5831	0.7805	0.5883	6.0709
WeSamBE [75]	0.7679	0.7955	0.7446	1.5105
Multimode Background [74]	0.7382	0.7389	0.7288	1.2614
SharedModel [137]	0.7503	0.8098	0.7474	1.4996
Spectral-360 [138]	0.7054	0.7345	0.6732	2.2722
Proposed	**0.9185**	**0.9197**	**0.9185**	**0.1177**

In this chapter, we have proposed two multi-scale deep learning architectures for moving object detection: modified ResNet-152 network with hybrid pyramidal pooling and multi-

scale contrast preserving deep learning architecture. The proposed algorithms results are evaluated qualitatively as well as quantitatively by comparing the results obtained by it with those of the different SOTA techniques that incorporate various deep learning and non-deep learning existing techniques. For empirical analysis, all these existing techniques are considered without altering the parameters. It may be noted that the proposed schemes surpasses, most of the existing state-of-the-art BGS techniques and also provides better accuracy.

To know the efficacy of the proposed algorithms, we have performed a quantitative comparison among them on five thermal sequences available at *changedetection.net* database are provided in Table 5.8. It may be found that the proposed modified ResNet-152 network with hybrid pyramidal pooling BGS scheme has provided a higher accuracy in terms of all considered measures against the proposed multi-scale contrast preserving deep learning architecture. Also, we have tested the proposed schemes on all the available videos (53 videos) available at *changedetection.net* and it may be observed from Table 5.9 that the proposed modified ResNet-152 network with hybrid pyramidal pooling scheme performance is better as compared to the proposed multi-scale contrast preserving deep learning architecture. Further, we have evaluated the proposed schemes on all the videos available at *TU-VDN* database are presented in Table 5.10. It may be found that the proposed modified ResNet-152 network with hybrid pyramidal pooling scheme attained better efficiency as compared to the multi-scale contrast preserving deep learning architecture.

Table 5.8: Quantitative comparisons on 5 thermal sequences of changedetection.net database

Quantitative measurements	Modified ResNet-152 network with hybrid pyramidal pooling	Multi-scale contrast preserving deep learning architecture
Avg.Precision	**0.9958**	0.9718
Avg.Recall	**0.9912**	0.9858
Avg.F-Measure	**0.9935**	0.9787
Avg.PWC	**0.0564**	0.5485

The proposed modified ResNet-152 network with a hybrid pyramidal pooling scheme and contrast preserving deep learning architecture yields better results for the test sequences. However, the proposed modified ResNet-152 network with hybrid pyramidal pooling BGS technique provides marginal outcomes if small moving objects are in excessive dynamism scenes. In such a scenario, spatial and temporal inter-dependency among

Table 5.9: Quantitative comparisons on all the sequences of changedetection.net database

Quantitative measurements	modified ResNet-152 network with hybrid pyramidal pooling	multi-scale contrast preserving deep learning architecture
Avg.Precision	**0.9905**	0.9185
Avg.Recall	**0.9853**	0.9197
Avg.F-Measure	**0.9878**	0.9185
Avg.PWC	**0.0343**	0.1177

Table 5.10: Quantitative comparisons on all the sequences of TU-VDN database

	Dust			Fog			Rain			Low Light		
Proposed algorithms	Avg. F-Measure	Avg. MCC	Avg. ACC	Avg. F-Measure	Avg. MCC	Avg. ACC	Avg. F-Measure	Avg. MCC	Avg. ACC	Avg. F-Measure	Avg. MCC	Avg. ACC
Modified ResNet-152 network with hybrid pyramidal pooling	**0.9131**	**0.9093**	**0.9926**	**0.8706**	**0.8745**	**0.9937**	**0.8792**	**0.8817**	**0.9959**	**0.8681**	**0.8709**	**0.9919**
Multi-scale contrast preserving deep learning architecture	0.8349	0.8286	0.9864	0.8589	0.8631	0.9929	0.8057	0.8036	0.9934	0.8554	0.8548	0.9908

the video frames can be used to identify small moving objects precisely. In the proposed multi-scale contrast preserving deep learning architecture, it may be observed that the use of a max-polling layer keeps the detail with maximum activation while discarding the information in different elements in a pooling area. In such a case, ordinal pooling may produce adequate results by considering all the elements in the pooling area with learning weights.

5.4 Conclusions

In this chapter, two multi-scale deep learning architectures for moving object detection are addressed for local change detection. In the proposed modified ResNet-152 network with hybrid pyramidal pooling scheme, a modified ResNet-152 network is induced on the multi-scale features extraction (MFE) block which is a hybridization of pyramidal pooling architecture (PPA) and various atrous convolutional layers for moving object detection. The encoder is configured using a modified ResNet-152 network to extract deep features from the video scenes where an image in RGB space is projected to a high dimensional feature space. In this work, we have proposed multi-scale features extraction (MFE) block integrated with the encoder network to enhance the feature learning capabilities that preserve sparse and dense deep features from challenging scenarios. We have explored PPA in the MFE block, which is a suitable contextual prior. Finally, we have proposed an adequate decoder network where up-sampling is performed to project the deep multi-scale features space to image-frame space. Again, in the multi-scale contrast preserving

deep learning architecture, a novel encoder network that deeply learned and extracted the sparse and dense features from the thermal input sequences. The proposed multi-scale contrast preservation (MSCP) block can precisely retain the contrast details of the in-depth features and act as a better feature representation block. The decoder network effectively classifies each pixel of the target scene to either foreground or background. For both the proposed algorithms, an end-to-end training mechanism is adapted to train the model, and a few input-ground-truth pair samples are used for the same. The proposed schemes are provided better-segmented results without utilizing any pre-or post-processing strategy.

The performance of the proposed algorithms is tested on benchmark databases: *changedetection.net*, and *Tripura University Video Dataset at Night Time (TU-VDN)* consisting of various outdoor and indoor image sequences. The proposed techniques are found to be robust against several real-life challenges, including irregular shades, higher dynamism in the background, camera jitter, thermal and illumination variations. The results obtained by the proposed modified ResNet-152 network with hybrid pyramidal pooling technique is validated against thirty-one state-of-the-art techniques, and the efficacy of the proposed multi-scale contrast preserving deep learning architecture is corroborated against twenty-eight existing techniques and found to be effective. To confirm our findings, we have used qualitative as well as quantitative analysis. It can be noted that the proposed schemes can detect the shape of the moving objects accurately and give higher values of evaluation measures for most of the considered experiments.

Chapter 6

Conclusions and Future Works

6.1 Conclusions

This thesis presents few new algorithms of image fusion to enhance the visual contents and few background subtraction scheme to detect the moving objects in the thermal video scene captured in a challenging indoor and outdoor scenarios. The challenging conditions considered are: low resolution (or) missing information, lack of structure such as shape and textural information, dynamic background, heat reflection from the surface, adverse weather conditions, etc. In this regard, the following contributory works have been proposed.

- In Chapter 2, we have proposed two contrast preservation with intensity variation approach for pixel level image fusion: fuzzy edge preserving intensity variation approach and weighted combination of maximum and minimum value selection strategy. In the proposed fuzzy edge preserving intensity variation approach, we have analyzed the spatial inter-dependency among the visible and thermal images to generate the salient feature map with reduced artifacts. However, the salient feature map cannot preserve sufficient edge details. Therefore the concept of the fuzzy edge is investigated in the visible image to obtain its edge. The fused image is generated by combining the salient feature map and edges of the visible image. Again in the proposed weighted combination of maximum and minimum value selection strategy, the detail feature map is generated using a maximum selection strategy in the source images that provide details of the object. However, the detail feature map cannot

preserve the subtle details from the source images. Therefore, the minimum selection strategy is used between the source images to produce the intermediate feature map with subtle details. A weighted-average fusion approach is explored to fuse the detailed and intermediate features. Both the proposed techniques produce the fused image with significant contrast and required details. The proposed schemes are tested on *TNO* benchmark database. The efficacy of the proposed fuzzy edge preserving intensity variation approach is validated against eight state-of-the-art schemes, and the efficiency of the proposed weighted combination of maximum and minimum value selection strategy is corroborated against the seven existing SOTA techniques. The performance of the proposed techniques is validated qualitatively and quantitatively in order to justify our findings. It is found that the proposed schemes provide better results compared to the state-of-the-art techniques.

- In Chapter 3, we have proposed two multi-scale features with deep learning architectures for feature level image fusion. In the first instance we proposed an integration of bi-dimensional empirical mode decomposition with two streams VGG-16 technique. The proposed bi-dimensional empirical mode decomposition (BEMD) strategy is integrated with a pre-trained VGG-16 network that can effectively handle the vagueness of infrared and visible images and can retain deep multi-layer features at different scales on the frequency domain. A fusion strategy is proposed to analyze the spatial inter-dependency between the deep features and preserve the complementary information from the source images precisely. Further, we proposed a non-subsampled contourlet transform induced two streams ResNet-50 network algorithm for image fusion. The non-subsampled contourlet transform (NSCT) geometrically transforms both the visual and the thermal images to get a shift-invariant, multi-direction, and multi-scale decomposition output. Two streams of parallel ResNet-50 networks: one for the low frequency and another for the high-frequency components of the NSCT are used here. A weighted combination strategy is proposed here to fuse the information from both visual and thermal image features output using the spatial inter-dependency among the pixels and precisely retain the correlative details from both the source images. Both the proposed techniques are found to be propagated lesser artifacts with rich edge details into the fused image. The efficiency of the proposed schemes is evaluated on the benchmark *TNO* database. The effi-

cacy of the proposed integration of bi-dimensional empirical mode decomposition with two streams VGG-16 scheme is corroborated against fifteen existing fusion schemes. Also, the performance of the proposed non-subsampled contourlet transform induced two streams ResNet-50 network algorithm is demonstrated against ten existing fusion schemes. We have used qualitative and quantitative analysis to ensure our findings and is found to be effective.

- In Chapter 4, we have proposed a kernel induced possibilistic fuzzy associate background subtraction to detect the moving objects from video scene. The proposed scheme follows two stages: background training and foreground segmentation. In the background construction stage, each pixel is modeled using a possibilistic fuzzy cost function in kernel induced space. The use of the induced kernel function will project the low dimensional data into a higher dimensional space and the use of the possibilistic function will construct a robust background model based on the density of the data in the temporal direction avoiding the noisy and outlier points. The performance of the proposed scheme is tested on the database: *changedetection.net*. The effectiveness of the proposed scheme is evaluated on different performance evaluation measures. The investigation is corroborated by comparing the results against twenty-nine existing state-of-the-art techniques and is found to be better.

- In Chapter 5, we have proposed two multi-scale deep learning architectures for moving object detection: modified ResNet-152 network with hybrid pyramidal pooling and multi-scale contrast preserving deep learning architecture. In the proposed modified ResNet-152 network with hybrid pyramidal pooling BGS technique, a pre-trained modified ResNet-152 network is adhered to as an encoder with a transfer learning mechanism to preserve the in-depth features against variation in grey value in the video scene. We have proposed a multi-scale features extraction block, a hybridization of pyramidal pooling architecture, and various atrous convolutional layers to extract multi-scale and multi-dimensional features at various levels. We have also proposed an efficient decoder network that uses low-level features from the encoder network and high-level features from the multi-scale features extraction block and up-scale the essential features into image space. In the proposed multi-scale contrast preserving deep learning architecture, we have developed an encoder network with skip connection to retain spatial information by considering the dis-

tinct neighborhood pixels in-depth at various levels. The proposed multi-scale contrast preservation block can precisely retain the contrast details of the deep features and act as a better feature representation block. The decoder network projects the extracted features at different layers into pixel-level accurately. The performance of the proposed algorithms is tested on benchmark databases: *changedetection.net*, and *Tripura University Video Dataset at Night Time (TU-VDN)*. The effectiveness of the proposed modified ResNet-152 network with hybrid pyramidal pooling technique is validated against thirty-one state-of-the-art techniques, and the efficacy of the proposed multi-scale contrast preserving deep learning architecture is corroborated against twenty-eight existing SOTA techniques and is found to be effective. To confirm our findings, we have used qualitative and quantitative analysis.

6.2 Future Works

The proposed techniques for thermal video surveillance are adequately able to address the image fusion and object detection tasks. However there are plenty of scopes in the field of thermal surveillance systems for establishing a robust automated surveillance system.

Thermal video processing plays a massive role in an automatic thermal surveillance system. The major drawback of most thermal video processing systems is their incapability to handle scenes with objects having the same temperature or the object's temperature same as the surface temperature. In such cases, there is substantial degradation of performance in terms of the separation of objects is observed. It may be noted that the proposed techniques are also unable to generate better results in such scenes. In this regard, some albedo analysis on visual images may be combined with the thermal image may improve the performance of the image fusion scheme.

One of the primary challenges in recently, deployed thermal surveillance system is developing distributed and collaborative sensing systems for providing the constituent sensors with the means to interpret each other's observations and measurements. The absence of global or even pairwise reference information effectively, isolates the individual sensors, leaving them unable to determine the meaning or relevance of other sensors' observations. While such reference information can be provided manually to systems comprised of a mere handful of sensors, systems deployed with hundreds and thousands

of sensors necessitate the development of automated approaches.

The common weakness of the proposed object detection techniques is their inability to handle densely crowded scenes. As the density of moving objects in the scene increases, a significant degradation in the performance in terms of object detection is observed. View variations and varying density of people as well as the ambiguous appearance of body parts, e.g. some parts of one object in the scene may be similar to another near-by object. This inability to deal with crowded scenes represents a significant problem. Solving such task is a very difficult task. Computing the tracks for multiple object with parallel computation is again more difficult. The future work may concentrate on developing some algorithms which may learns from a set of collective patterns of individuals from a specific testing scene. Some similar behaviors amongst crowd motion patterns combined with the features of the target candidates will be used for object tracking and behaviour analysis.

Bibliography

[1] K. P. Möllmann and M. Vollmer, *Infrared thermal imaging: fundamentals, research and applications*. Hoboken, NJ, USA: Wiley, 2017.

[2] A. Toet, "TNO image fusion dataset," *Figshare. data2014. https: // figshare.com/articles/TN_Image_ Fusion Dataset/1008029.*, 2014 (Last accessed on Dec.-16, 2019).

[3] A. Rogalski, *Infrared detectors*. New York: Gordon and Breach, 2000.

[4] R. Gade and T. B. Moeslund, "Thermal cameras and applications: a survey," *Machine Vision and Applications*, vol. 25, no. 1, pp. 245–262, 2014.

[5] R. Vadivambal and D. S. Jayas, "Applications of thermal imaging in agriculture and food industry-a review," *Food and Bioprocess Technology*, vol. 4, no. 2, pp. 186–199, 2011.

[6] L. Hoegner and U. Stilla, "Thermal leakage detection on building facades using infrared textures generated by mobile mapping," in *Proceedings of the IEEE Joint Urban Remote Sensing Event*, 2009, pp. 1–6.

[7] A. W. Lewis, S. T. Yuen, and A. J. Smith, "Detection of gas leakage from landfills using infrared thermography-applicability and limitations," *Waste Management & Research*, vol. 21, no. 5, pp. 436–447, 2003.

[8] G. Cong, D. Lu, Y. Lv, and Y. He, "A novel industrial safety IoTs architecture for external corrosion perception based on infrared," *Mobile Networks and Applications*, vol. 24, no. 4, pp. 1336–1345, 2019.

[9] J. M. Lloyd, *Thermal imaging systems*. Springer Science & Business Media, 2013.

[10] S. Li, X. Kang, L. Fang, J. Hu, and H. Yin, "Pixel-level image fusion: A survey of the state of the art," *Information Fusion*, vol. 33, pp. 100–112, 2017.

[11] J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: A survey," *Information Fusion*, vol. 45, pp. 153–178, 2019.

[12] T. T. Zin, H. Takahashi, T. Toriu, and H. Hama, "Fusion of infrared and visible images for robust person detection," *Image Fusion*, pp. 239–264, 2011.

[13] J. Ma, J. Zhao, and A. L. Yuille, "Non-rigid point set registration by preserving global and local structures," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 53–64, 2015.

[14] X. Zhang, P. Ye, S. Peng, J. Liu, K. Gong, and G. Xiao, "SiamFT: An RGB-infrared fusion tracking method via fully convolutional Siamese networks," *IEEE Access*, vol. 7, pp. 122 122–122 133, 2019.

[15] Y. Yan, J. Ren, H. Zhao, G. Sun, Z. Wang, J. Zheng, S. Marshall, and J. Soraghan, "Cognitive fusion of thermal and visible imagery for effective detection and tracking of pedestrians in videos," *Cognitive Computation*, vol. 10, no. 1, pp. 94–104, 2018.

[16] R. Singh, M. Vatsa, and A. Noore, "Integrated multilevel image fusion and match score fusion of visible and infrared face images for robust face recognition," *Pattern Recognition*, vol. 41, no. 3, pp. 880–893, 2008.

[17] P. Kumar, A. Mittal, and P. Kumar, "Fusion of thermal infrared and visible spectrum video for robust surveillance," in *Computer Vision, Graphics and Image Processing*, 2006, pp. 528–539.

[18] G. He, J. Ji, D. Dong, J. Wang, and J. Fan, "Infrared and visible image fusion method by using hybrid representation learning," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 11, pp. 1796–1800, 2019.

[19] X. Jin, Q. Jiang, S. Yao, D. Zhou, R. Nie, J. Hai, and K. He, "A survey of infrared and visual image fusion methods," *Infrared Physics & Technology*, vol. 85, pp. 478–501, 2017.

[20] G. Piella, "A general framework for multiresolution image fusion: From pixels to regions," *Information Fusion*, vol. 4, no. 4, pp. 259–280, 2003.

[21] J. Saeedi and K. Faez, "Infrared and visible image fusion using Fuzzy logic and population-based optimization," *Applied Soft Computing*, vol. 12, no. 3, pp. 1041–1054, 2012.

[22] E. Fendri, R. R. Boukhriss, and M. Hammami, "Fusion of thermal infrared and visible spectra for robust moving object detection," *Pattern Analysis and Applications*, vol. 20, no. 4, pp. 907–926, 2017.

[23] A. V. Vanmali and V. M. Gadre, "Visible and NIR image fusion using weight-map-guided Laplacian–Gaussian pyramid for improving scene visibility," *Sādhanā*, vol. 42, no. 7, pp. 1063–1082, 2017.

[24] Y. Liu, S. Liu, and Z. Wang, "A general framework for image fusion based on multi-scale transform and sparse representation," *Information Fusion*, vol. 24, pp. 147–164, 2015.

[25] H. Xu, Y. Wang, Y. Wu, and Y. Qian, "Infrared and multi-type images fusion algorithm based on contrast pyramid transform," *Infrared Physics & Technology*, vol. 78, pp. 133–146, 2016.

[26] P. Jagalingam and A. V. Hegde, "Pixel level image fusion: A review on various techniques," in *Proceedings of the 3rd World Conference on Applied Sciences, Engineering and Technology*, 2014, pp. 1–8.

[27] P. Hill, M. E. Al-Mualla, and D. Bull, "Perceptual image fusion using wavelets," *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1076–1088, 2016.

[28] F. Nencini, A. Garzelli, S. Baronti, and L. Alparone, "Remote sensing image fusion using the curvelet transform," *Information Fusion*, vol. 8, no. 2, pp. 143–156, 2007.

[29] Q. Zhang and B. l. Guo, "Multifocus image fusion using the nonsubsampled contourlet transform," *Signal Processing*, vol. 89, no. 7, pp. 1334–1346, 2009.

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

[31] D. P. Bavirisetti and R. Dhuli, "Two-scale image fusion of visible and infrared images using saliency detection," *Infrared Physics & Technology*, vol. 76, pp. 52–64, 2016.

[32] V. S. Petrovic and C. S. Xydeas, "Gradient-based multiresolution image fusion," *IEEE Transactions on Image Processing*, vol. 13, no. 2, pp. 228–237, 2004.

[33] J. J. Zong and T. S. Qiu, "Medical image fusion based on sparse representation of classified image patches," *Biomedical Signal Processing and Control*, vol. 34, pp. 195–205, 2017.

[34] 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.

[35] ——, "Medical image fusion via convolutional sparsity based morphological component analysis," *IEEE Signal Processing Letters*, vol. 26, no. 3, pp. 485–489, 2019.

[36] Y. Liu, X. Chen, J. Cheng, H. Peng, and Z. Wang, "Infrared and visible image fusion with convolutional neural networks," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 16, no. 3, pp. 1–20, 2018.

[37] Y. Liu, X. Chen, H. Peng, and Z. Wang, "Multi-focus image fusion with a deep convolutional neural network," *Information Fusion*, vol. 36, pp. 191–207, 2017.

[38] K. R. Prabhakar, V. S. Srikar, and R. V. Babu, "Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs." in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4724–4732.

[39] H. Li and X. Wu, "Densefuse: A fusion approach to infrared and visible images," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2614–2623, 2019.

[40] H. Li and X. J. Wu, "Infrared and visible image fusion using latent low-rank representation," *arXiv preprint arXiv:1804.08992*, pp. 1–6, 2018.

[41] H. Li, X. J. Wu, and J. Kittler, "RFN-Nest: An end-to-end residual fusion network for infrared and visible images," *Information Fusion*, vol. 73, pp. 72–86, 2021.

[42] C. Gao, D. Qi, Y. Zhang, C. Song, and Y. Yu, "Infrared and visible image fusion method based on ResNet in a nonsubsampled contourlet transform domain," *IEEE Access*, vol. 9, pp. 91 883–91 895, 2021.

[43] A. J. Rashidi and M. H. Ghassemian, "A new approach for multi-system/sensor decision fusion based on joint measures," *International Journal of Information Acquisition*, vol. 1, no. 02, pp. 109–120, 2004.

[44] V. E. Neagoe, A. D. Ropot, and A. C. Mugioiu, "Real time face recognition using decision fusion of neural classifiers in the visible and thermal infrared spectrum," in *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance*, 2007, pp. 301–306.

[45] Y. Zhao, Y. Yin, and D. Fu, "Decision-level fusion of infrared and visible images for face recognition," in *Proceedings of the IEEE Chinese Control and Decision Conference*, 2008, pp. 2411–2414.

[46] Y. Wang, W. Chen, and S. Mao, "Multi-sensor decision level image fusion based on Fuzzy theory and unsupervised FCM," in *Proceedings of the Remote Sensing of the Environment: 15th National Symposium on Remote Sensing*, vol. 6200, 2006, pp. 124–130.

[47] A. Wang, J. Jiang, and H. Zhang, "Multi-sensor image decision level fusion detection algorithm based on DS evidence theory," in *Proceedings of the IEEE International Conference on Instrumentation and Measurement, Computer, Communication and Control*, 2014, pp. 620–623.

[48] A. Vagale, A. Ņikitenko, E. Slava, and O. L. Osen, "Target identification using sensors of different nature," *Applied Computer Systems*, vol. 22, no. 1, pp. 28–35, 2017.

[49] H. Li, X. J. Wu, and T. S. Durrani, "Infrared and visible image fusion with resnet and zero-phase component analysis," *Infrared Physics & Technology*, vol. 102, pp. 1–22, 2019.

[50] M. Haghighat and M. A. Razian, "Fast-FMI: non-reference image fusion metric," in *Proceedings of the 8th IEEE International Conference on Application of Information and Communication Technologies*, 2014, pp. 1–3.

[51] B. K. S. Kumar, "Multifocus and multispectral image fusion based on pixel significance using discrete cosine harmonic wavelet transform," *Signal, Image and Video Processing*, vol. 7, no. 6, pp. 1125–1143, 2013.

[52] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[53] F. Sattar, L. Floreby, G. Salomonsson, and B. Lovstrom, "Image enhancement based on a nonlinear multiscale method," *IEEE Transactions on Image Processing*, vol. 6, no. 6, pp. 888–895, 1997.

[54] M. Cristani, M. Farenzena, D. Bloisi, and V. Murino, "Background subtraction for automated multisensor surveillance: A comprehensive review," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, no. 343057, pp. 1–24, 2010.

[55] C. Stauffer and W. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 747–757, 2000.

[56] A. Elgammal, R. Duraiswami, D. Harwood, and L. S. Davis, "Background and foreground modeling using nonparametric kernel density estimation for visual surveillance," *Proceedings of the IEEE*, vol. 90, no. 7, pp. 1151–1163, 2002.

[57] X. Liu, G. Zhao, J. Yao, and C. Qi, "Background subtraction based on low-rank and structured sparse decomposition," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2502–2514, 2015.

[58] T. Bouwmans, "Background subtraction for visual surveillance: A fuzzy approach," in *Handbook on Soft Computing for Video Surveillance*, S. K. Pal, A. Petrosino, and L. Maddalena, Eds. New York: Chapman and Hall/CRC, 2012, p. 36.

[59] T. Bouwmans, S. Javed, M. Sultana, and S. K. Jung, "Deep neural network concepts for background subtraction: A systematic review and comparative evaluation," *Neural Networks*, vol. 117, pp. 8–66, 2019.

[60] B. Bhanu and J. Han, "Kinematic-based human motion analysis in infrared sequences," in *Proceedings of the 6th IEEE Workshop on Applications of Computer Vision*, 2002, pp. 208–212.

[61] F. Xu, X. Liu, and K. Fujimura, "Pedestrian detection and tracking with night vision," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 1, pp. 63–71, 2005.

[62] T. Elguebaly and N. Bouguila, "Finite asymmetric generalized Gaussian mixture models learning for infrared object detection," *Computer Vision and Image Understanding*, vol. 117, no. 12, pp. 1659–1671, 2013.

[63] B. N. Subudhi, S. Ghosh, and A. Ghosh, "Change detection for moving object segmentation with robust background construction under Wronskian framework," *Machine Vision and Applications*, vol. 24, no. 4, pp. 795–809, 2013.

[64] D. K. Rout, B. N. Subudhi, T. Veerakumar, and S. Chaudhury, "Spatio-contextual Gaussian mixture model for local change detection in underwater video," *Expert Systems with Applications*, vol. 97, pp. 117 – 136, 2018.

[65] L. Maddalena and A. Petrosino, "A self-organizing approach to background subtraction for visual surveillance applications," *IEEE Transactions on Image Processing*, vol. 17, no. 7, pp. 1168–1177, 2008.

[66] K. Makantasis, A. Nikitakis, A. D. Doulamis, N. D. Doulamis, and I. Papaefstathiou, "Data-driven background subtraction algorithm for in-camera acceleration in thermal imagery," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 9, pp. 2090–2104, 2018.

[67] K. Kim, T. H. Chalidabhongse, D. Harwood, and L. Davis, "Real-time foreground–background segmentation using codebook model," *Real-time Imaging*, vol. 11, no. 3, pp. 172–185, 2005.

[68] M. Heikkila and M. Pietikainen, "A texture-based method for modeling the background and detecting moving objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 657 –662, 2006.

[69] J. W. Davis and V. Sharma, "Background-subtraction in thermal imagery using contour saliency," *International Journal of Computer Vision*, vol. 71, pp. 161–181, 2007.

[70] O. Barnich and M. Van Droogenbroeck, "ViBe: A universal background subtraction algorithm for video sequences," *IEEE Transactions on Image Processing*, vol. 20, no. 6, pp. 1709–1724, 2011.

[71] T. S. Haines and T. Xiang, "Background subtraction with Dirichlet process mixture models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 4, pp. 670–683, 2014.

[72] P. L. St-Charles, G. A. Bilodeau, and R. Bergevin, "SuBSENSE: A universal change detection method with local adaptive sensitivity," *IEEE Transactions on Image Processing*, vol. 24, no. 1, pp. 359–373, 2015.

[73] B. N. Subudhi, S. Ghosh, S. C. Shiu, and A. Ghosh, "Statistical feature bag based background subtraction for local change detection," *Information Sciences*, vol. 366, pp. 31–47, 2016.

[74] H. Sajid and S. C. S. Cheung, "Universal multimode background subtraction," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3249–3260, 2017.

[75] S. Jiang and X. Lu, "WeSamBE: A weight-sample-based method for background subtraction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 9, pp. 2105–2115, 2017.

[76] C. L. P. Chen, H. Li, Y. Wei, T. Xia, and Y. Y. Tang, "A local contrast method for small infrared target detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 1, pp. 574–581, 2014.

[77] J. Han, Y. Ma, J. Huang, X. Mei, and J. Ma, "An infrared small target detecting algorithm based on human visual system," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 3, pp. 452–456, 2016.

[78] A. Singha and M. K. Bhowmik, "Salient features for moving object detection in adverse weather conditions during night time," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 10, pp. 3317–3331, 2020.

[79] C. Guyon, T. Bouwmans, and E. Zahzah, "Robust principal component analysis for background subtraction: Systematic evaluation and comparative analysis," *INTECH, Principal Component Analysis*, vol. 10, pp. 223–238, 2012.

[80] J. Seo and S. D. Kim, "Recursive on-line $(2D)^2PCA$ and its application to long-term background subtraction," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2333–2344, 2014.

[81] S. E. Ebadi, V. G. Ones, and E. Izquierdo, "Efficient background subtraction with low-rank and sparse matrix decomposition," in *Proceedings of the IEEE International Conference on Image Processing*, 2015, pp. 4863–4867.

[82] W. Cao, Y. Wang, J. Sun, D. Meng, C. Yang, A. Cichocki, and Z. Xu, "Total variation regularized tensor RPCA for background subtraction from compressive measurements," *IEEE Transactions on Image Processing*, vol. 25, no. 9, pp. 4075–4090, 2016.

[83] C. Li, X. Wang, L. Zhang, J. Tang, H. Wu, and L. Lin, "Weighted low-rank decomposition for robust grayscale-thermal foreground detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 4, pp. 725–738, 2017.

[84] L. X. Wu, X., "Adaptive pixel-block based background subtraction using low-rank and block-sparse matrix decomposition," *Multimedia Tools and Applications*, vol. 78, pp. 16 507 – 16 526, 2019.

[85] M. H. Sigari, N. Mozayani, and H. R. Pourreza, "Fuzzy running average and Fuzzy background subtraction: Concepts and application," *International Journal of Computer Science and Network Security*, vol. 8, no. 2, pp. 138–143, 2008.

[86] H. Zhang and D. Xu, "Fusing color and texture features for background model," in *Proceedings of the 3rd International Conference on Fuzzy Systems and Knowledge Discovery*, 2006, pp. 887–893.

[87] P. Chiranjeevi and S. Sengupta, "Neighborhood supported model level Fuzzy aggregation for moving object segmentation," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 645–657, 2014.

[88] T. Bouwmans and F. El-Baf, "Modeling of dynamic backgrounds by type-2 Fuzzy Gaussians mixture models," *MASAUM Journal of Basic and Applied Sciences*, vol. 1, no. 2, pp. 265–277, 2009.

[89] F. El-Baf, T. Bouwmans, and B. Vachon, "Fuzzy integral for moving object detection," in *Proceedings of the IEEE International Conference on Fuzzy Systems*, 2008, pp. 1729–1736.

[90] Z. Zhao, T. Bouwmans, X. Zhang, and Y. Fang, "A Fuzzy background modeling approach for motion detection in dynamic backgrounds," in *Proceedings of the International Conference on Multimedia and Signal Processing*, 2012, pp. 177–185.

[91] L. Maddalena and A. Petrosino, "A fuzzy spatial coherence-based approach to background/foreground separation for moving object detection," *Neural Computing and Applications*, vol. 19, no. 2, pp. 179–186, 2010.

[92] W. Kim and C. Kim, "Background subtraction for dynamic texture scenes using Fuzzy color histograms," *IEEE Signal Processing Letters*, vol. 19, no. 3, pp. 127–130, 2012.

[93] M. I. Chacon-Murguia and S. Gonzalez-Duarte, "An adaptive Neural-Fuzzy approach for object detection in dynamic backgrounds for surveillance systems," *IEEE Transactions on Industrial Electronics*, vol. 59, no. 8, pp. 3286–3298, 2012.

[94] Z. L. Li, W. M. Liu, and Y. Zhang, "A background modeling method based on adaptive Fuzzy estimation," *Journal of South China University of Technology*, vol. 41, pp. 77–81, 2013.

[95] R. Rajkumar, K. Sukkiramathi, and R. Vijayanandh, "Fuzzy C-means (FCM) clustering and adaptive network-based Fuzzy inference system (ANFIS) for dynamic background subtraction," *Solid State Technology*, vol. 63, no. 6, pp. 3974–3988, 2020.

[96] M. Balcilar and A. C. Sonmez, "Region based Fuzzy background subtraction using Choquet integral," in *Adaptive and Natural Computing Algorithms*, M. Tomassini, A. Antonioni, F. Daolio, and P. Buesser, Eds., 2013, pp. 287–296.

[97] Y. L. Qiao, K. L. Yuan, C. Y. Song, and X. Z. Xiang, "Detection of moving objects with Fuzzy color coherence vector," *Mathematical Problems in Engineering*, vol. 2014, no. 2, pp. 1–8, 2014.

[98] T. Yu, J. Yang, and W. Lu, "Dynamic background subtraction using histograms based on Fuzzy C-means clustering and Fuzzy nearness degree," *IEEE Access*, vol. 7, pp. 14 671–14 679, 2019.

[99] B. N. Subudhi, T. Veerakumar, S. Esakkirajan, and A. Ghosh, "Kernelized fuzzy modal variation for local change detection from video scenes," *IEEE Transactions on Multimedia*, vol. 22, no. 4, pp. 912–920, 2020.

[100] M. Braham and M. Van Droogenbroeck, "Deep background subtraction with scene-specific convolutional neural networks," in *Proceedings of the International Conference on Systems, Signals and Image Processing*, 2016, pp. 1–4.

[101] M. Babaee, D. T. Dinh, and G. Rigoll, "A deep convolutional neural network for video sequence background subtraction," *Pattern Recognition*, vol. 76, pp. 635–649, 2018.

[102] R. Wang, F. Bunyak, G. Seetharaman, and K. Palaniappan, "Static and moving object detection using flux tensor with split Gaussian models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 414–418.

[103] Y. Wang, Z. Luo, and P. M. Jodoin, "Interactive deep learning method for segmenting moving objects," *Pattern Recognition Letters*, vol. 96, pp. 66–75, 2017.

[104] T. P. Nguyen, C. C. Pham, S. V. Ha, and J. W. Jeon, "Change detection by training a triplet network for motion feature extraction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 2, pp. 433–446, 2019.

[105] Y. Yan, H. Zhao, F. J. Kao, V. M. Vargas, S. Zhao, and J. Ren, "Deep background subtraction of thermal and visible imagery for pedestrian detection in videos," in *International Conference on Brain Inspired Cognitive Systems*, 2018, pp. 75–84.

[106] Z. Hu, T. Turki, N. Phan, and J. T. Wang, "A 3D Atrous convolutional long short-term memory network for background subtraction," *IEEE Access*, vol. 6, pp. 43 450–43 459, 2018.

[107] Y. Wang, L. Zhu, and Z. Yu, "Foreground detection for infrared videos with multiscale 3-D fully convolutional network," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 5, pp. 712–716, 2019.

[108] O. Tezcan, P. Ishwar, and J. Konrad, "BSUV-Net: A fully-convolutional neural network for background subtraction of unseen videos," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 2774–2783.

[109] L. A. Lim and H. Y. Keles, "Learning multi-scale features for foreground segmentation," *Pattern Analysis and Applications*, vol. 23, no. 3, pp. 1369–1380, 2020.

[110] M. C. Bakkay, H. A. Rashwan, H. Salmane, L. Khoudour, D. Puig, and Y. Ruichek, "BSCGAN: Deep background subtraction with conditional generative adversarial networks," in *Proceedings of the 25th IEEE International Conference on Image Processing*, 2018, pp. 4018–4022.

[111] D. Sakkos, E. S. Ho, and H. P. Shum, "Illumination-Aware multi-task GANs for foreground segmentation," *IEEE Access*, vol. 7, pp. 10 976–10 986, 2019.

[112] W. Zheng, K. Wang, and F. Y. Wang, "A novel background subtraction algorithm based on parallel vision and Bayesian GANs," *Neurocomputing*, vol. 394, pp. 178–200, 2020.

[113] Y. Wang, P. M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar, "CDnet 2014: An expanded change detection benchmark dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 387–394.

[114] N. Goyette, P. M. Jodoin, F. Porikli, J. Konrad, and P. Ishwar, "Changedetection.net: A new change detection benchmark dataset," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2012, pp. 1–8.

[115] S. K. Pal, R. King *et al.*, "Image enhancement using smoothing with Fuzzy sets," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 11, no. 7, pp. 494–500, 1981.

[116] B. K. S. Kumar, "Image fusion based on pixel significance using cross bilateral filter," *Signal, Image and Video Processing*, vol. 9, no. 5, pp. 1193–1204, 2015.

[117] J. Ma, Z. Zhou, B. Wang, and H. Zong, "Infrared and visible image fusion based on visual saliency map and weighted least square optimization," *Infrared Physics & Technology*, vol. 82, pp. 8–17, 2017.

[118] J. C. Nunes, Y. Bouaoune, E. Delechelle, O. Niang, and P. Bunel, "Image analysis by bidimensional empirical mode decomposition," *Image and Vision Computing*, vol. 21, no. 12, pp. 1019–1026, 2003.

[119] T. Veerakumar, B. N. Subudhi, and S. Esakkirajan, "Empirical mode decomposition and adaptive bilateral filter approach for impulse noise removal," *Expert Systems with Applications*, vol. 121, pp. 18–27, 2019.

[120] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London. Series A: mathematical, physical and engineering sciences*, vol. 454, no. 1971, pp. 903–995, 1998.

[121] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, pp. 1–14, 2015.

[122] H. Li, Y. Chai, and Z. Li, "Multi-focus image fusion based on nonsubsampled contourlet transform and focused regions detection," *Optik*, vol. 124, no. 1, pp. 40–51, 2013.

[123] A. 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.

[124] X. Wang, L. Yin, M. Gao, Z. Wang, J. Shen, and G. Zou, "Denoising method for passive photon counting images based on block-matching 3D filter and non-subsampled contourlet transform," *Sensors*, vol. 19, no. 11, pp. 1–15, 2019.

[125] M. Habibzadeh, M. Jannesari, Z. Rezaei, H. Baharvand, and M. Totonchi, "Automatic white blood cell classification using pre-trained deep learning models: ResNet and inception," in *Proceedings of the 10th International Conference on Machine Vision*, vol. 10696, 2018, pp. 1–10.

[126] O. Wichrowska, N. Maheswaranathan, M. W. Hoffman, S. G. Colmenarejo, M. Denil, N. de Freitas, and J. Sohl-Dickstein, "Learned optimizers that scale and generalize," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2017, pp. 3751–3760.

[127] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[128] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[129] H. Li, X. J. Wu, and J. Kittler, "Infrared and visible image fusion using a deep learning framework," in *Proceedings of the 24th International Conference on Pattern Recognition*, 2018, pp. 2705–2710.

[130] C. Liu, Y. Qi, and W. Ding, "Infrared and visible image fusion method based on saliency detection in sparse domain," *Infrared Physics & Technology*, vol. 83, pp. 94–102, 2017.

[131] 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, pp. 11–26, 2019.

[132] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, and L. Zhang, "IFCNN: A general image fusion framework based on convolutional neural network," *Information Fusion*, vol. 54, pp. 99–118, 2020.

[133] M. Filippone, F. Masulli, and S. Rovetta, "Applying the possibilistic C-means algorithm in kernel-induced spaces," *IEEE Transactions on Fuzzy Systems*, vol. 18, no. 3, pp. 572–584, 2010.

[134] A. Chan, V. Mahadevan, and N. Vasconcelos, "Generalized Stauffer and Grimson background subtraction for dynamic scenes," *Machine Vision and Applications*, vol. 22, pp. 751–766, 2011.

[135] F. Seidel, C. Hage, and M. Kleinsteuber, "pROST - a smoothed Lp-norm robust online subspace tracking method for realtime background subtraction in video," *Machine Vision and Applications*, vol. 25, no. 5, pp. 1227–1240, 2014.

[136] P. St-Charles, G. Bilodeau, and R. Bergevin, "Universal background subtraction using word consensus models," *IEEE Transactions on Image Processing*, vol. 25, no. 10, pp. 4768–4781, 2016.

[137] Y. Chen, J. Wang, and H. Lu, "Learning sharable models for robust background subtraction," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2015, pp. 1–6.

[138] M. Sedky, M. Moniri, and C. C. Chibelushi, "Spectral-360: A physics-based technique for change detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 399–402.

[139] V. M. Mondéjar-Guerra, J. Rouco, J. Novo, and M. Ortega, "An end-to-end deep learning approach for simultaneous background modeling and subtraction," in *Proceedings of the 30th British Machine Vision Conference*, 2019, pp. 1–12.

[140] S. Lee, G. Lee, J. Yoo, and S. Kwon, "WisenetMD: Motion detection using dynamic background region analysis," *Symmetry*, vol. 11, no. 5, pp. 1–15, 2019.

[141] S. Bianco, G. Ciocca, and R. Schettini, "Combination of video change detection algorithms by genetic programming," *IEEE Transactions on Evolutionary Computation*, vol. 21, no. 6, pp. 914–928, 2017.

[142] M. Braham, S. Piérard, and M. Van Droogenbroeck, "Semantic background subtraction," in *Proceedings of the IEEE International Conference on Image Processing*, 2017, pp. 4552–4556.

[143] M. O. Tezcan, P. Ishwar, and J. Konrad, "BSUV-net 2.0: Spatio-temporal data augmentations for video-agnostic supervised background subtraction," *IEEE Access*, vol. 9, pp. 53 849–53 860, 2021.

[144] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the International Conference on Machine Learning*, 2015, pp. 448–456.

[145] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2881–2890.

[146] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, Atrous convolution, and fully connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[147] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, pp. 1–6, 2016.

[148] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 648–656.

[149] L. A. Lim and H. Y. Keles, "Foreground segmentation using convolutional neural networks for multiscale feature encoding," *Pattern Recognition Letters*, vol. 112, pp. 256–262, 2018.

[150] P. L. St-Charles and G. A. Bilodeau, "Improving background subtraction using local binary similarity patterns," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2014, pp. 509–515.

[151] M. Hofmann, P. Tiefenbacher, and G. Rigoll, "Background segmentation with feedback: The pixel-based adaptive segmenter," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2012, pp. 38–43.

[152] Y. Goyat, T. Chateau, L. Malaterre, and L. Trassoudaine, "Vehicle trajectories evaluation by static video sensors," in *Proceedings of the Intelligent Transportation Systems Conference*, 2006, pp. 864–869.

[153] B. Wang and P. Dudek, "A fast self-tuning background subtraction algorithm," in *Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 395–398.

[154] L. Unzueta, M. Nieto, A. Cortés, J. Barandiaran, O. Otaegui, and P. Sánchez, "Adaptive multicue background subtraction for robust vehicle counting and classification," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 2, pp. 527–540, 2012.

[155] F. C. Cheng, S. C. Huang, and S. J. Ruan, "Illumination-sensitive background modeling approach for accurate moving object detection," *IEEE Transactions on Broadcasting*, vol. 57, no. 4, pp. 794–801, 2011.

[156] J. E. Ha and W. Lee, "Foreground objects detection using multiple difference images," *Optical Engineering*, vol. 49, no. 4, pp. 047 201–1–047 201–5, 2010.

[157] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, 1999, pp. 246–252.

[158] N. M. Oliver, B. Rosario, and A. P. Pentland, "A Bayesian computer vision system for modeling human interactions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 831–843, 2000.

[159] M. Wu and X. Peng, "Spatio-temporal context for codebook-based dynamic background subtraction," *AEU-International Journal of Electronics and Communications*, vol. 64, no. 8, pp. 739–747, 2010.

List of publications

The thesis is based on the following papers/preprints.

1. M. K. Panda, B. N. Subudhi, T. Veerakumar, and M. S. Gaur, "Edge preserving image fusion using intensity variation approach," *in Proceedings of the IEEE REGION 10 CONFERENCE (TENCON)*, 2020, pp. 251–256.

2. M. K. Panda, B. N. Subudhi, T. Veerakumar, and M. S. Gaur, "Pixel-level visual and thermal images fusion using maximum and minimum value selection strategy," *in Proceedings of the IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security*, 2020, pp. 1–6.

3. Badri Narayan Subudhi, Manoj Kumar Panda,Thangaraj Veerakumar, Vinit Jakhetiya, and Sankaralingam Esakkirajan, "Kernel-induced possibilistic fuzzy associate background subtraction for video scene," *IEEE Transactions on Computational Social Systems*, pp. 1-12, 2022.

4. M. K. Panda, B. N. Subudhi, T. Veerakumar and, Vinit Jakhetiya, "Two streams ResNet-50 network for infrared and visible image fusion," *Expert Systems With Applications* (under review).

5. M. K. Panda, B. N. Subudhi, T. Veerakumar and, Vinit Jakhetiya, "Integration of bi-dimensional empirical mode decomposition with two streams deep learning network for infrared and visible image fusion," *in Proceedings of the IEEE International Conference on European Signal Processing Conference (EUSIPCO)* (under review).

6. M. K. Panda, T. Veerakumar, B. N. Subudhi, and Vinit Jakhetiya, "Bayesian's probabilistic strategy for feature fusion from visible and infrared Images," *The Visual Computer* (under review).

7. M. K. Panda, B. N. Subudhi, Thierry Bouwmans, Vinit Jakhetiya, and T. Veerakumar, "An encoder-decoder network with multi-scale pulling for local change detection," *in Proceedings of the IEEE International Conference in Image Processing* (under review).

8. M. K. Panda, Akhilesh Sharma, Vatsalya Bajpai, B. N. Subudhi, T. Veerakumar and Vinit Jakhetiya "Encoder and decoder network with ResNet-50 and global average feature pooling for local change detection," *Computer Vision and Image Understanding* (under review).