A Study on Multispectral Image and Video Fusion with some Applications

Thesis submitted by

GANGAPURE VIJAY NARAYAN

Doctor of Philosophy (*Engineering*)

Department of Electronics and Telecommunication Engineering

FACULTY COUNCIL OF ENGINEERING AND TECHNOLOGY

JADAVPUR UNIVERSITY

KOLKATA, INDIA

2017

JADAVPUR UNIVERSITY

KOLKATA- 700 032, INDIA

INDEX NO. 295/14/E

1. **Title of the thesis:**

   "A Study on Multispectral Image and Video Fusion with some Applications"

2. **Name, Designation & Institution of the Supervisor:**

Dr. Ananda Shankar Chowdhury
Associate Professor
Department of Electronics and Telecommunication Engineering
Jadavpur University, Kolkata- 700 032, India

3. **List of publication:**

(In International Journals)

(i) **Vijay N. Gangapure**, Sudipta Banerjee, Ananda S. Chowdhury: Steerable local frequency based multispectral multifocus image fusion. Information Fusion, Elsevier, 23: 99-115 (2015)

(ii) **Vijay N. Gangapure**, S. Nanda, A.S. Chowdhury: Superpixel based Causal Multisensor Video Fusion, IEEE Transactions on Circuits and Systems for Video Technology (2017). [in press, DOI: 10.1109/TCSVT.2017.2662743]

(In International Conferences)

(iii) **Vijay N. Gangapure**, Susmit Nanda, Ananda S. Chowdhury, Xiaoyi Jiang: "Causal Video Segmentation Using Superseeds and Graph Matching", Graph based Representation in Pattern Recognition (GbRPR 2015), 10^{th} IAPR International Workshop, Springer LNCS 9069, Beijing China, May 13-15 (2015), 282-291.

(iv) Sudipta Banerjee, **Vijay N. Gangapure**, Ananda S. Chowdhury: "Multispectral Multifocus Image Fusion with Guided Steerable Frequency and Improved Saliency", 9^{th} Indian Conference Conference on Computer Vision, Graphics and Image Processing (ICVGIP 2014), IISc Bangalore India, December 14-18 (2014), 9:1-9:8.

(v) **Vijay N. Gangapure**, R. Sarkar, A.S. Chowdhury: "2.5D Palmprint Recognition using Signal level Fusion and Graph based Matching", Ninth International Conference on Advances in Pattern Recognition (ICAPR), ISI Bangalore, India, December 27-30 (2017). (accepted)

4. **List of patents: NIL**

5. **List of Presentations in National/International:**

(i) Presented a paper entitled “Multispectral Multifocus Image Fusion with Guided Steerable Frequency and Improved Saliency,Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP 2014)"’, December 14-18, 2014, IISc Bangalore, India.

(ii) Presented a paper entitled “Causal Video Segmentation Using Superseeds and Graph Matching"’, 10th IAPR-TC15 Workshop on Graph-based Representations in Pattern Recognition (GbR 2015), May 13-15, 2014, NLPR, Institute of Automation. Chinese Academy of Sciences (CASIA) , Beijing, China.

(iii) To be presented a paper entitled "2.5D Palmprint Recognition using Signal level Fusion and Graph based Matching", Ninth International Conference on Advances in Pattern Recognition (ICAPR), December 27-30, 2017, ISI Bangalore, India.

CERTIFICATE

This is to certify that the thesis entitled ***"A Study on Multispectral Image and Video Fusion with some Applications"*** *submitted by* ***Shri. Gangapure Vijay Narayan****, who got his name registered on* ***21/04/2014 [D-7/E/331/14]*** *for award of Ph.D. (Engg.) degree of Jadavpur University is absolutely based upon his own work under the supervision of* ***Dr. Ananda Shankar Chowdhury****, and that neither his thesis nor any part of the thesis has been submitted for any degree/ diploma or any other academic award anywhere before.*

Dr. Ananda Shankar Chowdhury

Signature of the Supervisor and date with office seal

Abstract

In this thesis, we address different challenges in multispectral image and video fusion. Our first problem is on multispectral multifocus image fusion. A novel focus measure, based on steerable local frequency,is proposed.The proposed measure is shown to perform uniformly well across different spectra like visible, near infra-red and thermal. We further enhance the quality of the multispectral multifocus fusion by guided filtering and a graph based saliency model. The second problem is on causal multispectral video fusion. In this connection, we first solve the problem of causal video segmentation. An efficient causal video segmentation method is proposed using superseeds and graph matching. We then design a novel superpixel based causal multispectral video fusion method suitable for real-time surveillance applications. As a part of this solution, superpixel based spatio-temporal saliency model as well as superpixel based multiple fusion rules are developed. For the third problem, we consider an application of fusion in the domain of multi-biometric recognition. Here, we perform a signal level fusion of 2D and 3Dpalmprint data and apply a graph based recognition strategy. The proposed solution is shown to yield high recognition accuracy.

Acknowledgements

First and foremost I would like to express my special appreciation and thanks to my advisor Dr. Ananda Shankar Chowdhury. He has been a tremendous mentor for me. I appreciate all his contributions of time, ideas, motivation and support to make my Ph.D. experience productive and stimulating. The joy and enthusiasm he has for his research was contagious and motivational for me, even during tough times in the Ph.D. pursuit. I am also thankful for the excellent example he has provided as a successful researcher and very good human being.

I gratefully acknowledge Prof. P. Venkateswaran, Head of the Department of Electronics and Telecommunication Engineering, Jadavpur University, all the faculty members, technical staff and administrative staff members of the department for their care and warmth shown towards me during my research period in this University. Their active support and encouragement always helped me during my difficult situations. Especially, I am grateful to Prof. Subir K. Sarkar, Prof. Amit Konar, Prof. K. K. Mallik, and Prof. Iti Saha Misra for their support and encouragement during the research. I am also thankful to the Faculty of Engineering and Technology (FET), Jadavpur University and Research sections for their guidance and assistance.

The members of the Imaging, Vision and Pattern Recognition (IVPR) group have contributed immensely to my personal and professional time at Jadavpur University. The group has been a source of friendships as well as good advice and collaboration. I would like to acknowledge group members Sanjay Kumar Kuanar, Gautam Bhattacharya. We worked together and I very much appreciate their enthusiasm, intensity, willingness to help me in my research. I really enjoyed working with PG students Sudipta Banerjee, Susmit Nanda and Rahul Sarkar. I really appreciate their efforts during the work. The other past and present group members that I have had the pleasure to work with or alongside of are Arindam Sikdar, Rukmini Roy, Rameshwar Panda, Dheeraj Zha, Kunal Bhushan Ranga, Amit Kumar Sagar, Soman Chakraborty, Prasenjit Mudi, Ranodip Das and all fellow researchers in the department.

I would also like to show my gratitude to the Department of Technical Education, Maharashtra State, Mantralaya, Mumbai; Directorate of Technical Education, Mumbai; AICTE, New Delhi for providing financial and moral support for my Ph.D work. I would also like to say thank to my honorable higher authorities and colleagues without whom's support and constant motivation I would not have been able to carry out this work.

Finally, but by no means least, I am enormously grateful to my family for almost unbelievable support. They are the most important people in my world and I dedicate this thesis to them. My parents Late Sri. Narayan Gangapure, and Smt. Sharda Gangapure; wife Madhura and sons Atharva & Arnav who believed in me and supported me. Their love and care filled me with strength and optimism and kept me going. If it were not for the sacrifices made by my family, I would not be here today. I dedicate this work to them. I would also like to express my sincere gratitude to my brothers, sisters, friends, family members and my in-laws for their constant encouragement and support rendered during my course of research study.

Jadavpur University
Kolkata- 700 032

(Gangapure Vijay Narayan)

Dedicated to …

My Parents

My Wife

And

My Sons

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter provides an outline of multispectral multifocus image and multispectral video fusion with a fusion application in multimodal biometry. In section 1.1, we take the overview of the research work. In section 1.2 we discuss the motivation behind the work undertaken. Section 1.3 presents the key contributions. In section 1.4, we provide an overview of how the rest of this thesis is organized.

1.1 Overview

Image and video fusion remains a challenging problem in the area of image/video analysis and computer vision. It has widespread applications in diverse fields like medicine, surveillance, military and law enforcement, remote sensing, biometrics, manufacturing, intelligent robots [1–5]. The main objective of any image/video fusion algorithm is to design highly accurate and computationally efficient strategy to combine information from two or a more source images/frames of a scene to produce more informative fused image. The fused image/frame can be further exploited for extraction of useful information. The image and video fusion can also be referred to as static and dynamic image fusion.

General requirements of image or video fusion algorithms are as follows [6].

1. The relevant information from source images to be fused should not be lost in the fused one.
2. Fusion process itself should not introduce any kind of artifacts, inconsistencies that may distract the purpose.
3. The fused image sequence should be temporally stable and consistent.

4. Fusion process should be shift and rotation invariant.

Depending on the application requirement we can choose the level of information representation at which fusion process actually take place. Sorted in ascending order of abstraction we have- *Signal or pixel level fusion* (low level), *Feature level fusion* (mid level) and *Symbolic or Decision level fusion* (high level). Actual image fusion process may occur in spatial or transform domain [7].

Images or videos to be fused may not be aligned or may have some relative translation, rotation, scale, and other geometric transformations in relation to each other, or not on the same coordinate system. So image to image registration is necessary by which we transform such images into one coordinate system or align with respect to each other. In image fusion it is presumed that the images to be fused are registered.

1.1.1 Image fusion

The source images to be fused can be captured under differently varying conditions. Accordingly, we have different types of fusion schemes like- *Multiview fusion*, *Multitemporal fusion*, *Multifocus fusion*, *Multispectral fusion*, *Multimodal fusion*.

- **Multiview image fusion:** The objective of multiview image fusion is to fuse images of same modality of a scene captured at the same time but from the different places.This is to just supply complementary information from different view points to produce more informative fused image of a scene, see Fig. 1.1.

FIGURE 1.1: Multiview fusion: Example.

- **Multitemporal image fusion:** In this, images of the same scene usually of the same modality but taken at different times are fused together to detect changes. Fig. 1.2 shows such an example of digital subtraction angiography in medicine.

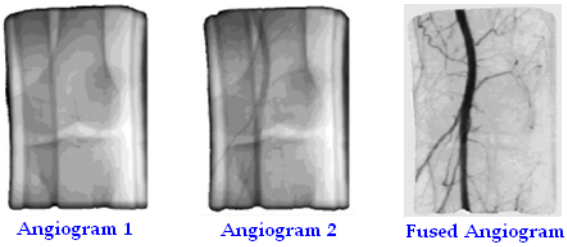


FIGURE 1.2: Multitemporal fusion: Example.

- **Multifocus image fusion:** Addresses image quality degradation problem in image acquisition system. The overall goal is to bring all the objects in the scene in focus and the resultant image is called all-in-focus image. Please refer to Fig. 1.3 for an example.

FIGURE 1.3: Multifocus fusion: Example.

- **Multimodal/Multispectral image fusion:** Here, images of different modalities e.g. visible, infrared, thermal spectra; CT, MRI, PET. are fused together to decrease the amount of data to emphasis band specific information in the fused video. Fig. 1.4 shows an example of concealed weapon detection by fusion of visible and infrared spectrum images of scene.

1.1.2 Video fusion

Recently, the image fusion methods are extended to video fusion. Video fusion have been become very popular for various applications like video surveillance, super resolution reconstruction, video snapshots or summarization generation, restoration

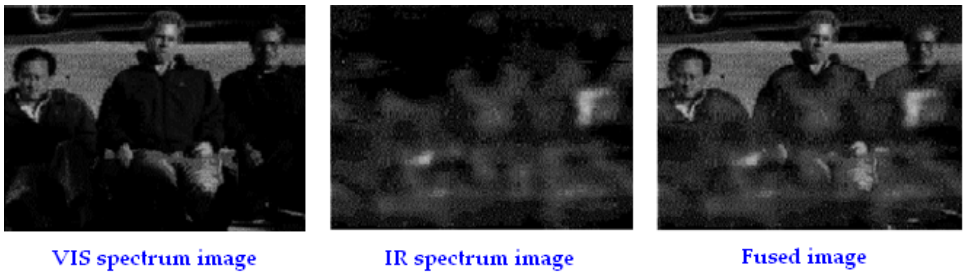


FIGURE 1.4: Multimodal/Multispectral fusion: Example.

enhancement. Depending on the number of videos sensors involved in the fusion, we can classify video fusion in two categories- the *Intra video fusion* or *Inter video fusion*. In case of intra-video fusion the subsequent frames from a single video are processed using fusion framework to produce enhanced video frame or snapshot or summarization. While in case of inter-video fusion, multispectral videos of a scene captured by more than one video sensors are involved. These videos are fused together to produce more informative video of a scene.

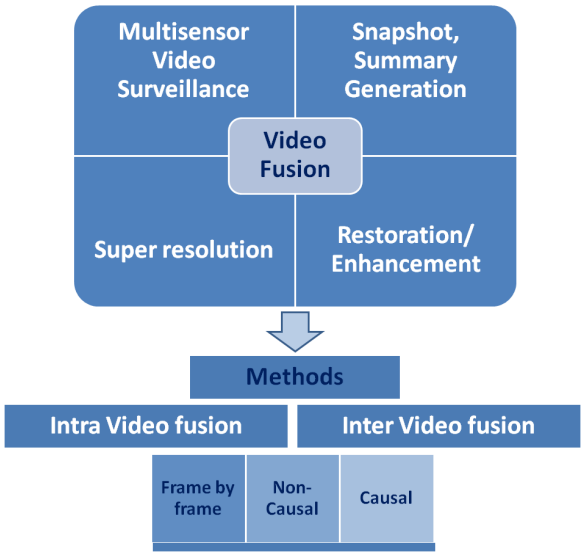


FIGURE 1.5: Video fusion: Applications and Methods scenario

The advancement in multispectral video sensors technology and it's availability at low cost have motivated researchers to come up with novel multispectral multisensor video surveillance systems. Such systems are proved to be very effective and have

outperformed the single sensor based systems in uncontrolled and adverse conditions like low llumination, shadowing, smoke, dust, unstable background, and camouflage. The fusion plays a very important role in such multisensor video surveillance systems. Here, the fusion can be employed to combine information from the multispectral videos of a scene; from different spectra like visible (VIS), infrared (IR) or thermal infrared (TH-IR); to produce more informative fused video. The produced fused video can describe a scene more accurately and precisely as compared to any of the individual modalities. This helps in analysis of further important surveillance tasks like anomalous event detection [8] and person re-identification [9].

Early video fusion methods, which are just extension of static image fusion algorithms [3, 4, 10, 11], suffer from temporal instability and inconsistencies. Some later methods [12–14] addressed this problem by utilizing the information from adjacent past and future frames. Depending on the utilizing information from either from past or/and future frames, we can classify these methods as- *Non-causal* (Utilize both past and future frames) and *Causal* (Utilize only past frames). For real time surveillance applications the future frames may not be available at the time of processing of current frame.

1.1.3 Some applications of fusion

Fusion provides solution to a wide variety of applications such as:

1) *Biomedical imaging*: Some examples are- Fusing X-ray computed tomography (CT) and magnetic resonance (MR) images, computer assisted surgery, spatial registration of 3-D surface.

2) *Biometric authentication*: Multimodal biometric systems involving multiple biometric traits like fingerprint, iris, face, 2D and 3D palmprint .

3) *Manufacturing*: Finds application in electronic circuit and component inspection, product surface measurement and inspection, non-destructive material inspection,

manufacture process monitoring, complex machine/device diagnostics and intelligent robots on assembly lines.

4) *Military and law enforcement*: For detection, tracking, identification of ocean (air, ground) target or event, concealed weapon detection, battle-field monitoring and night pilot guidance.

5) *Remote sensing*: Using various parts of the electro-magnetic spectrum Sensors-from black-and-white aerial photography to multi-spectral active microwave space-borne imaging radar.

6) *Intelligent robots*: Require motion control, based on feedback from the environment from visual, tactile, force/torque, and other types of sensors, stereo camera fusion, intelligent viewing control, automatic target recognition and tracking.
Out of the above applications, we focus on multimodal biometrics in the present thesis. A general overview of such a multimodal biometric system is presented below.

Fusion can play very important role in biometric authentication systems. Basically there are two major approaches for biometric authentication- *unimodal biometric systems* and *multimodal biometrics systems* (Multibiometrics). Very recently, multibiometric approaches have been become very popular due to its significant advantages over unimodal systems, like improvement in recognition performance, improving population coverage, deterring spoof attacks, increasing the degrees of freedom, and reducing the failure-to-enroll rate [15]. For illustration of multibiometric system, refer to Fig. 1.6 The key to successful multi-biometric system is in an effective fusion scheme, which is necessary to combine the information presented by multiple traits. Fusion for the multi-biometric system is relatively new area. The information from the multiple traits can be integrated at several different levels and we can subdivide them into two main categories- *prior to matching fusion* and *after matching fusion*. In prior to matching fusion case, the information integration takes place before matching. This category again subdivided into- Sensor level fusion (low level) and feature level fusion (mid level). In case of after matching fusion, the information integration takes place after matching. Again this category is subdivided into- Match score level,

Rank level and Decision level fusion (high level)[16–18]. Biometric systems that integrate information at an early stage of processing are believed to be more effective than those systems which perform integration at a later stage [19]. On the contrary, feature level fusion have also provided better recognition results but suffers from difficulties to achieve in practice due to unknown relationship between feature spaces of different modalities and curse of dimensionality. Fusion at decision level is too rigid since only limited amount of information is available at this level. Integration at matching score level is normally preferred in many past systems due to ease of access and combining matching scores. Thus, innovative combination of fusion could be used to achieve robust performance.

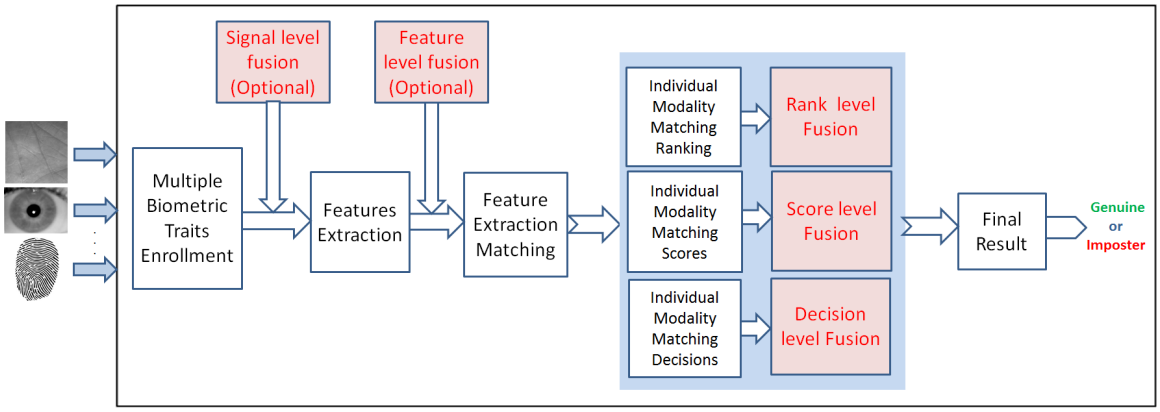


FIGURE 1.6: Generic Multimodal biometric system.

Today, the multbiometric systems are in developing stage. A very few multibiometric systems are practically deployed and in its enrollment stage viz. UIDAI's (Unique Identification Authority of India) Aadhaar project, by Government of India, to provide identification to each resident. The growing security concerns as well as spoofing attack threats clearly suggests the future prospects of the multbiometric systems. Most importantly, the effective way in which fusion of the multiple traits takes place, ultimately decides the performance of these systems. Recent literature shows the use of various kinds of physiological, medico-chemical, behavioral and soft biometrics (attribute or characteristics or traits) combinations being used in multbiometric systems. The use of physiological biometrics from various body regions like face,

hand, ocular are mostly preferred due to their inherent benefits like universality, distinctiveness, invariance, collectability, performance, acceptability and circumvention [20]. The use of fingerprint, iris, and face have been become very popular in unimodal systems and are being used with other novel traits in multbiometrics systems, like 2D and 3D palmprint, hand-geometry, hand vein pattern, finger knuckle point, ear shape, teeth, tongue print, retina. Recently use of 2D and 3D palmprints have become very popular due to its unique complementary nature and benefits. The combination of these two traits could be very effective to avoid spoofing attacks.

1.2 Motivation

Multifocus image fusion is a category of fusion that involves images at different focus levels. The limited depth of field in cameras does not enable capturing of all objects in a scene at a consistently maintained focus level. Therefore, to obtain a highly focused image, information from a sequence of defocused images needs to be integrated together. Focus measure quantifies the amount of information present in the images. It can be used to determine whether information from a typical source image is relevant or not [21, 22]. An image having high focus measure contains significant information and plays a crucial role in image fusion. So, the effectiveness of the focus measure to a large extent determines the quality of image fusion. The focus measure should be robust enough to implement intra-spectral fusion effectively. Majority of the existing multifocus image fusion methods are in the visual spectrum only. Very less work is thus far reported in other spectra like near-infrared (NIR) , Thermal infrared (TH-IR).

We observed that the performance of many recently proposed focus measures [23] is variant to the spectral content of the source images to be fused viz. VIS, NIR and TH-IR. To alleviate this problem one could employ spectrum dependent focus measures to achieve acceptable performance. This prompted us to the formulation

of our first problem to work on. So, the objective is to design a novel multispectral focus measure which performs uniformly and effectively across different spectra and can objectively determine the focus level of images. Furthermore, the proposed focus measure should be robust enough to implement intra-spectral fusion effectively.

Majority of inter multispectral video fusion approaches are non-causal and are transform domain based [12–14, 24, 25]. The transform domain approaches, in spite of some inherent advantages over spatial domain, suffers from approximation issues in implementation and may be quite expensive if employed at higher scales to achieve robust performance. On the other hand spatial domain approaches are more accurate and precise, but suffers from computational inefficiency and thus limiting its employability for real time applications. To address these problems, in our second work we propose a solution to causal multispectral video fusion. We present a novel superpixel based framework for causal multispectral videos fusion (CMVF) in spatial domain. The proposed approach can be very useful for real time video surveillance applications.

Currently 2D and 3D palmprint biometrics has become popular due to its advantages like- high distinctiveness, robustness, and high user-friendliness. The 2D palmprint acquisition is less robust against illumination changes; contamination on palm can substantially affect recognition rate; can be easily copied and counterfeited, so vulnerable to spoofing attacks [26]. We can overcome the challenges 2D palmprint recognition is facing by the use of 3D palmprint. It is also more difficult to fake/copy/and counterfeit 3D palmprint as compared to 2D palmprint to spoof the biometric system [27]. Also, 3D palmprint is always preferred than other 3D biometric technologies like 3D face and 3D ear [28, 29] due to its certain benefits. Compared to 3D face, 3D palmprint is not affected by challenges associated with various facial expressions and is much easier to acquire and more user friendly than 3D ear. To build more robust and highly secure palmprint recognition system we can take advantage of the highly discriminative texture rich information from 2D palmprint

and depth information from 3D palmprint. Biometric systems that integrate information at an early stage of processing are believed to be more effective than those systems which perform integration at a later stage[30]. In literature there are very few attempts to fuse the 2D and 3D palmprint information in recognition [26, 31–33]. Majority of these approaches are based either on feature or score level fusion. As per best of our knowledge, up to date there is no work reported which exploited signal/sensor level fusion of 2D and 3D palmprint data. So, in our third work we plan to propose a multimodal biometric system based on fusion of 2D, 3D palmprint. The fused 2.5D palmprint representation can be more informative than the original individual 2D and 3D palmprints and can be useful in improvement in recognition performance.

1.3 Contributions

We now state the key contributions of this thesis work.

1. The **first problem** we addressed is on multispectral multifocus image fusion. The key contributions of this work are as follows.

    - We propose new focus measure based on steerable local frequency (SLF) based interest point detection. For this we suggest the construction of the oriented analytic image. The proposed focus measure captures all the sharp features in different orientations and hence perform well across different spectra (VIS, NIR, TH).

    - We employ the proposed focus measure for intra multispectral multifocus image fusion. The proposed fusion scheme achieves improved fusion performance across different spectra.

    Further, to achieve better fusion performance we propose a solution based on

guided steerable local frequency and improved Saliency (GSLF-IS). The contributions of this solution are as follows.

- Judicious application of the guided edge preserving filtering in two phases. In the first phase, the source images to be fused are enhanced using guided filtering keeping the source images same as the guidance images. In the second phase, the steerable local frequency maps of the enhanced source images are further refined using guided filtering. In this case, the enhanced source images are used as the guidance images.
- Development of an improved model of saliency based on Graph-based visual saliency (GBVS), Spectral residual saliency (SRA) and Laplacian saliency (LS).
- The improved saliency map is combined with the guided steerable local frequency map to generate good fusion results across all spectra.

2. Our **second problem** is on multisensor causal video fusion from different spectra. To provide solution, we first worked on causal video segmentation. For this we present a solution using superseeds and local graph matching. The major contributions of this part of work are:
   - A novel method of label propagation based on graph matching.
   - Use of superseeds for achieving better segmentation.
   - Unlike some of the existing approaches, we do not use any post-processing steps to achieve superior segmentation performance.

   Based on this work, next, we propose a superpixel based causal multisensor video fusion (CMVF) algorithm. Visible and infrared video pairs are fused using this algorithm to obtain highly accurate information in a time-efficient manner. Here, the main contributions are:

- New superpixel level spatio-temporal saliency models.
- Novel multiple fusion rules for saliency based grouping of superpixels.
- Low execution time making it amenable for real-time surveillance applications.

Comprehensive comparison with several existing approaches on a number of publicly available datasets clearly indicate the advantage of our video fusion method.

3. The **third problem** is an application of fusion for multi-biometric recognition. We propose a novel multimodal biometric system based on fusion of aligned 2D and 3D palmprints. The main contributions of this work are as follows.
   - Signal level fusion of 2D and 3D palmprints with local standard deviation based novel fusion rule to produce more informative 2.5D palmprint data.
   - Use of graph based template generation and matching framework.

   The Comprehensive comparison with very recent score and feature level fusion based approaches shows the superiority of the proposed system.

1.4 Thesis Organization

The rest of the thesis is organized in the following manner:

In Chapter 2, in first part, we discuss the steerable local frequency (GSLF) based solution for multispectral focus measure and its application n multifocus image fusion. In the second part we present extension to enhance performance. Here, innovative use of guided filtering and improved saliency model is discussed. The detailed experimentation analysis with datasets and performance metrics are also given in subsequent sections.

In chapter 3, we provide a novel superpixel based framework for casual multispectral/multisensor video fusion. We divide the discussion into two parts. In the first part, we provide a novel causal video segmentation solution using superseeds and graph matching for semantic segmentation of the input visible spectrum video. In the second part, we provide very details of the proposed multispectral video fusion. The efficient superpixel based spatial and temporal saliency detection models and fusion rules to achieve final fusion are discussed in details. The chapter ends with very comprehensive experimentation analysis and comparisons. The details of used datasets and performance metrics used can also be found.

In chapter 4, we provide a solution to multimodal biometric system based on 2D, 3D palmprints. We took detail overview of biometric systems based on 2D, 3D and 2D+3D palmprints. In the next parts we provide the details of the proposed framework. First the preprocessing of 2D and 3D palmprints and signal level fusion to produce 2.5D palmprint data is discussed. Next, graph based template generation and matching is discussed in detail. Finally, the chapter ends with experimental results section containing details about dataset used, performance metrics , comparisons and analysis.

Finally, Chapter 5 concludes the work presented in this thesis, and provides some directions for future research.

Chapter 2

Multispectral Multifocus Image Fusion

The objective of multifocus image fusion is to integrate source images of a scene captured at different focal lengths to produce all-in-focus image. The primary step in any multifocus image fusion to determine the focus quality of source images to be fused. In this chapter, first, we propose a novel steerable local frequency based focus measure which is robust to spectral content of source images to be fused. Secondly, based on this focus measure we propose an efficient multispectral multifocus image fusion scheme. Further, we enhance the fusion performance by introducing the innovative use of improved saliency model and guided filtering. The proposed approaches fulfills the need of uniform multispectral focus measure and as well as of an efficient multispectral multifocus image fusion scheme.

2.1 Introduction

The multifocus image fusion which entails fusion of relevant information from two or more source images obtained at different focal points into a composite image of better quality (all in focus image). Imaging cameras, particularly those with long focal lengths, usually have only a finite depth of field. In an image captured by those cameras, only those objects within the depth of field of the camera are focused, while other objects are blurred. To obtain an image that is in focus everywhere, we need to fuse the images taken from the same view point under different focal settings. The aim of multifocus image fusion is to integrate complementary and redundant information from multiple images to create a composite result that contains a better

description of the scene than any of the individual source images [21, 22]. In order to determine, information from which of the source images needs to be integrated in the fused result, the notion of focus measure or image sharpness or image clarity comes into play. For example, a certain region in the source image A has higher focus measure than that of the same region in source image B. So, the fused result, F will contain information from A. Thus, a good focus measure inherently improves the fusion performance. So, fundamental step behind any multifocus fusion lies in the determination of the focus quality of the source images. Furthermore, the fusion scheme applied based on the focus quality measure determines the quality of the final fused image.

2.1.1 Focus measures

A focus measure is an objective function of digital images that gives a single value for each input image as the indicator of its focusing status or sharpness status. The desirable characteristics any focus measure should have are listed as follows [21, 22]:

- *Unimodality*: The focus measure should be unimodal i.e. it should have only one maxima corresponding to the highest level of focus. Unimodality ensures unambiguous solution during the search for best focused position in the image.
- *Monotonicity*: On either side of the maxima , a focus measure should be monotonic so that focus measure values should be different for different levels of defocus.
- *Defocus and Noise Sensitivity*: Ideally a focus measure should be sensitive to defocus and insensitive to noise.
- *Effective Range*: It is the defocus range over which a focus measure maintains its reasonable sensitivity. The broader its effective range, the better is the focus measure.

- *Computational Efficiency*: A focus measure should not be too computationally complex.
- *Variability*: A good focus measure should not vary dramatically from one case to another, that is, it should be "repeatable" over different target scenes and optical systems.
- Focus measure should be independent of image content.
- Focus measure should be independent of image modality.

Focus measures can be broadly classified in to two categories depending on the domain in which the focus measure is determined of source image. These are spatial domain and transform domain focus measures.

Spatial domain focus measures:
As name implies the focus measure is computed in the original spatial domain of the image under consideration. Spatial domain focus measures are further classified into four broad categories, such as, *derivative based*, *statistics based*, *histogram based* and *intuition based*.

The derivative based algorithms suppose that well-focused images have more high frequency content than the defocused images. It considers that the neighboring pixels in images with high frequency content have large differences in intensity. They apply convolution masks to an image to obtain the derivatives. The magnitude of the derivative vectors computed using norms yields the required focus measure value. However, in computing derivatives, these algorithms are highly sensitive to noise. A few examples of such focus measures include thresholded absolute gradient, squared gradient, Brenner gradient, Tenenbaum gradient, energy of Laplacian and sum modified laplacian. Thresholded absolute gradient [34] sums the absolute value of the first derivative that is larger than a certain threshold. Squared gradient [34] on the other hand sums the squared differences, making larger gradients exert more influence. Brenner gradient [35] computes the first difference between a pixel and its neighbor

with a horizontal/vertical distance of 2. Tenenbaum gradient, popularly known as Tenengrad [21, 36], convolves an image with Sobel operators; sums the square of the gradient vector components. Energy of laplacian (EOL) [37] convolve an image with one of the laplacian convolution masks to compute the second derivative. The final output is obtained as the sum of the squares of the convolution results along height and width of the image. Sum modified laplacian (SML) [38] computes the sum of the absolute values of the convolution of an image with laplacian operators.

The statistics based algorithms distinguish focused images from defocused images using variance and correlation. They are generally less sensitive to noise than derivative-based algorithms. Variance [36, 39] computes variations in gray level among image pixels. It uses the power function to amplify larger differences from the mean intensity instead of simply enhancing the high-intensity values. Normalized variance [36, 39] on the other hand normalizes the final output with the mean intensity, compensating for the differences in average image intensity among different images. Auto-correlation algorithm incorporates the auto-correlation among the different pixels [40, 41].

Histogram based algorithms use histograms $H(i)$ (i.e., the number of pixels with intensity i in an image) to analyze the distribution and frequency of image intensities. Range algorithm [42], a focus measure belonging to the above category computes the difference between the highest and the lowest intensity levels. Entropy algorithm [42] assumes that focused images contain more information than defocused images. Thus, image having higher entropy intrinsically has larger focus measure.

Intuitive based focus measures make use of heuristic approaches to derive the focus measure. Thresholded content [39, 43] sums the pixel intensities above a given threshold θ to compute the required focus measure. Similarly, thresholded pixel count [39] counts the number of pixels having intensity below a given threshold. Image Power [34] obtains the focus measure as the sum of the square of image intensities above a given threshold, θ. Spatial Frequency [44] (SF) is a focus measure

which can be considered as a derivative based focus measure due to its dependence on gradients. It indicates the overall activity level in an image $f(i, j)$. The gradients are computed by measuring the intensity differences between adjacent pixels along horizontal and vertical lines. The gradients are squared so as to enhance the contributions of the larger gradients more than those of the smaller gradients.

Transform domain focus measures:

The transform domain focus measures based on wavelet transforms are proposed such as W1, W2 and W3. Wavelet algorithm W1 [45, 46] uses the Daubechies D6 wavelet. The resultant decomposition of an image consists of four sub-images; LL, HL, LH, and HH belonging to low-low, high-low, low-high and high-high sub-bands. The algorithm sums the absolute values in the HL, LH, and HH regions. Wavelet algorithm W2 [45, 46] computes the sum of the variances in the HL, LH, and HH regions. The mean values in each region are computed from the absolute values. Wavelet algorithm W3[45, 46] is same as that of the previous version with the exception of the mean values in each region, which are computed without using absolute values. Recently proposed locally adaptive laplacian mixture model based focus measure algorithm [47] examines the statistics of the wavelet coefficients to evaluate the sharpness measurement. It has been established that the marginal distribution of the wavelet coefficients in the high frequency bands varies with images having different focus levels. To provide a quantitative measurement of the degree of blur in an image, a locally adaptive laplacian mixture model is used to formulate the marginal distribution of wavelet coefficients.

Interest point detector based focus measures

An interest point is a point in an image where certain property changes significantly. Probably the most frequently considered property is intensity. As a result, interest points can act as an important tool in defining the features present in an image, such as blobs, corners, edges. It can be observed that the number of detected interest

points decreases when the image is blurred/defocused. Thus, it can serve the purpose of determining the focus measure of an image i.e. the number of interest points detected in a well-focused image will be very higher as compared to other defocused images [23, 48]. Another important aspect of interest point detectors is that they exploit the concept of phase congruency, local frequency of an image without being dependent solely on intensity [49, 50].

2.1.2 Multifocus image fusion

A generic multifocus image fusion is illustrated in Fig. 2.1.

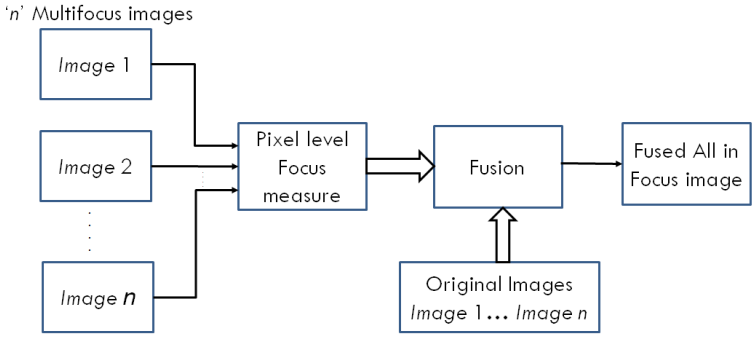


FIGURE 2.1: Generic Multifocus image fusion.

Many multifocus fusion algorithms are available in the literature. These algorithms operate at pixel level or region level, and in spatial as well as transform domains. Spatial domain pixel level algorithms are popular due to their computational efficiency [6]. Multiresolution transform based algorithms are preferred nowadays due to their robust performance [37]. Within the transform domain, Discrete Wavelet Transform (DWT) based algorithms, though perform better than the Laplacian Pyramid Transform (LPT), have limited orientation selectivity [51]. More improved multiresolution transform techniques include Stationary Wavelet Transform (SWT), Curvelet Transform (CVT), Contourlet Transform (CT), Dual Tree Complex Wavelet Transform (DTCWT) and Non-Subsampled Contourlet Transform (NSCT) [37, 51]. In [37], Li et al. has evaluated the performance of such multiresolution transforms for multifocus image fusion in the visual spectrum only. Benes et al. [52]

proposed a new multifocus image fusion algorithm for thermal images where they employed pixel level weighted averaging based on modified EOL. But such a linear combination often fails to preserve the original information in source images leading to degradation in the fusion performance.

Existing literature clearly suggests that designing a focus measure and applying it for fusion across different spectra still poses a considerable challenge. In this work, we propose a novel focus measure using steerable local frequency based interest point detection. A recent work is reported where pixel intensities at different orientations are considered for obtaining a focus measure [53]. However, phase of a pixel carry more useful information than intensity [54]. To the best of our knowledge, this phase information has not been captured in different orientations earlier. In this work, we make the local frequency of the pixels, the spatial derivative of the local phase, steerable to obtain a good focus measure. For this purpose, we suggest the construction of the oriented analytic image. The proposed focus measure captures all possible sharp features in different orientations and hence performs well across different spectra. Further, we employ our focus measure for multispectral multifocus image fusion. Detailed experimentation reveal much improved multifocus image fusion performances across different spectra.

We further enhance the performance of aforementioned multispectral multifocus fusion method by using novel improved saliency model and guided filtering. We use guided edge preserving filter in two phases. In the first phase source images to be fused are enhanced using guided filtering. In the second phase, the obtained SLF maps of these enhanced source images are further refined using guided filtering, resulting guided steerable local frequency maps (GSLF). Our improved model of saliency is based on Graph-based visual saliency (GBVS), Spectral residual saliency (SRA) and Laplacian saliency (LS). This improved saliency map is combined with the GSLF map to generate good fusion results across all spectra. The results shows improvement in fusion performance over our previous approach.

2.2 Related works

This section reviews the select major approaches described in the literature for focus measures and multifocus image fusion.

2.2.1 Focus measures

In [55], Liu et al. evaluated the performance of eighteen focus measures, e.g., EOG (Energy of Gradient), SML (Sum of Laplacian), EOL (Energy of Laplacian), TEN (Tenengrad) etc. for microscopic images from the different categories as discussed earlier, see section 2.5.2. All these focus measures are based on variations in pixel intensities only. These methods have several drawbacks, like, the performance variation with spectral content of the source images, insensitivity to defocus, fluctuation with noise content and narrow effective range.

To overcome these limitations, recently, Minhas et al. [53] proposed a novel efficient focus measure. The main objective was to build depth map generation for shape from focus (SFF) application. The proposed method uses steerable filters for depth map estimation from the sequence of images acquired at varying focus plane. Further, the highest gradient information is exploited at the desired orientation is utilized for the purpose.

In another method Tian and Chen [47] examined statistics of details wavelet coefficients to perform sharpness measurement in the input image. The marginal distribution of the wavelet coefficients is different for images with different focus levels. So, the degree of focus is measured by exploiting the marginal distribution of the wavelet coefficients.

Zhao et al. [56] related the degree of gray level surface curvature to the sharpness of image region. The new parameter, neighbor distance (ND), is proposed as a measure of pixel's sharpness/focus. First, the oriented distance(OD) is used to measure

the image surface curvature at a particular point. The sum of ODs along different directions surrounding the pixel in image is called Neighbor Distance(ND).

Aforementioned attempts to measure focus are limited to the visual spectrum only. In other spectra like thermal and near-infrared, less focus measure works are reported due to unavailability of scene viewing in case of manual focusing, and limited resolution as well as lack of auto-focus features in the cameras.

Faundez-Zanuy et al. [57] addressed the problem of determining the optimal focus position in the thermal images, for the first time. For this they used existing five focus measures such as EOG, TEN, EOL, SML, SF and Crete et al. [58]. The performance of these measures is analyzed in order to obtain most suitable focus measure for thermal images. Among these measures, EOG, EOL and SML offered good performance.

Very recently, detection of interest point based focus measure for multispectral images has gained popularity. Zukal et al. [48], introduced the determination of focus measure via interest point detection. They proposed focus measure based on three types of interest point detectors- FAST(Features from Accelerated Segmented Test), Fast Hessian(FH), Harris Laplace (HL). These measures evaluated against some of the standard focus measures such as EOG, SML, TEN, SF. The results show that these focus measures performs better than the standard focus measures for thermal images but lag behind in the visual and the near-infrared spectra. We observed that the all these measures are based on intensity [23].
However, some of the interest point detection methods are found to use frequency and phase congruency [49, 50].

In [52], Benes et al. proposed a new focus measure which is based on EOL (Energy of Laplacian). It is computed as product of average value of EOL in certain neighborhood multiplied by variance in the same neighborhood.

2.2.2 Multifocus image fusion

Many multifocus fusion algorithms have been proposed based on some of the above mentioned focus measures. These algorithms operate at pixel level or region level, and in spatial as well as transform domains. Multiresolution analysis is widely adapted technique to perform image fusion. Spatial domain pixel level algorithms are popular due to their computational efficiency [6]. Multiresolution transform based algorithms are preferred nowadays due to their robust performance [37]. But spatial domain processing techniques are normally preferred due to fewer operations and more suitable for real time applications [57].

Within the transform domain, Discrete Wavelet Transform (DWT) based algorithms, though perform better than the Laplacian Pyramid Transform (LPT), have limited orientation selectivity [51]. More improved multiresolution transform techniques include Stationary Wavelet Transform (SWT), Curvelet Transform (CVT), Contourlet Transform (CT), Dual Tree Complex Wavelet Transform (DTCWT) and Non-Subsampled Contourlet Transform (NSCT) [37, 51].
In [37], Li et al. has has given very comprehensive performance comparison of some multiresolution transforms (DWT, SWT, DTCWT, CVT, CT, NSCT) for multifocus image fusion, but within visual spectrum only. The multifocus image fusion performance of the DTCWT and NSCT methods partially better in terms of some of the metrics. While for infrared-visible and medical fusion applications NSCT performs better.

Tian and Chen [47] proposed a new wavelet based multifocus image fusion method. To perform sharpness/focus measurement, the statistics of wavelet coefficients are examined. In the fusion framework, the wavelet transform is applied on each input image. Then the detail and approximation wavelet coefficients are combined using different pixel level focus measure fusion rule in the transform domain. And finally the fused image is obtained by applying inverse wavelet transform on the fused wavelet coefficients. The proposed approach outperforms some of the earlier

wavelet transform based fusion methods but not tested in other spectra.

Zhao et al. [56] proposed multifocus image fusion scheme based on novel neighbor distance (ND) based focus measure computation. They choose the multi-resolution transform based fusion, due to disadvantage of block effect in construction of ND in spatial domain. For this purpose multiscale ND analysis framework based on ND filter is proposed. First obtain multiscale low frequency component image and ND image sequence. Then multiple sets of low frequency and ND components are combined together using choose maximum (CM), Salience/Match measure with threshold or Choose maximum with consistency check. The consistency verification is performed further but is optional. Finally, a fused image is produced by reconstruction using fused low frequency component and ND image sequence. The performance of proposed approach is evaluated against some standard transform domain methods like DWT, FSD (Filter Subtract Decimate hierarchical pyramid), GRP (Gradient Pyramid), RAP (Ratio of Low pass pyramid), LAP (Laplacian Pyramid), SVT (Support Value Transform), SWT and NSCT. The propose method have shown very good performance over all these methods, but applied in visible spectrum only.

Benes et al. in [52], proposed a first of its kind a novel approach for multifocus image fusion for thermal images. The algorithm involves three major steps- Measurement of activity level, selection of best images for fusion, and combination of the selected images. In the first step, modified EOL based pixel level activity level (as measure of focus) measurement is performed on each image in the set. in the second step, the fixed suitable number of images are selected surrounding the images having peaks for the activity level. Finally, in the third step, selected images are combined using simple proposed activity level based pixel weighted average rule rule. The superiority of the proposed approach for thermal multifocus image fusion is shown intra-comparisons and by highlighting the reduction in error in the temperature measurement. But such a linear combination often fails to preserve the original information in source images and may lead to a degradation in the fusion

performance.

2.3 Theoretical foundations

In this section, we provide the theoretical foundations behind the proposed method. In particular, analytic image, steerable filters, image saliency models and guided filtering are discussed in details.

2.3.1 Analytic image

The Fourier transform or its variant produces global, phase and magnitude information, which lacks spatial localized information. So it can not be used to process non-stationery signals like images for better localized information representation. One solution is to use spatial domain localized frequency analysis schemes which has become an important and powerful tool in signal representation. This gives us local magnitude and phase information. As Oppenheim had demonstrated through a series of experiments that the signal phase serves an important role even more than intensity [54]. The phase conveys more information regarding signal structure than magnitude, especially in case of images. It is also highly immune to noise and contrast distortions which are desirable features in image processing. We can employ the analytic signal representation as it provides an easy solution for recovering local phase from the signal [49].

Let us introduce the concept of analytic signal in 1D which can be extended to higher dimensions [49]. Given a time domain signal s(t) in 1D, its analytic signal is defined as: The local phase of signal can be computed by representing it analytically and is known as *analytic signal*.

$$s_A(t) = s(t) - j s_H(t) \tag{2.1}$$

where, $s_H(t)$ is the Hilbert transform of $s_A(t)$. An image can be treated as a $2D$ spatial domain signal. Corresponding *analytic image* can be expressed as:

$$I_A(x, y) = I(x, y) - jI_H(x, y) \tag{2.2}$$

where, $I_H(x, y)$ is the Hilbert transformation of $I(x, y)$. Argument of $I_A(x, y)$, defined in the spatial domain, is referred to as the local phase of $I(x, y)$. Khan et al. [49] have used the local frequency of an image to capture the dominant regions in an image (a dominant region will contain many pixels with high local frequencies). The local frequency can be determined easily as it is the spatial derivative of local phase. High value of local frequency at a particular pixel of an image indicates the presence of interest point at that location. Concept of quadrature pair of filters can be introduced in this context. Quadrature pair of filters has same frequency response but differ in phase by an angle of 90^o, i.e., in effect they must be Hilbert transforms of each other [59].

2.3.2 Steerable Gaussian filter

Steerable filters can be defined as a special class of filters in which an arbitrarily oriented filter can be designed using a linear combination of a set of basis filters [60, 61]. The directional derivative of a 2D Gaussian function is steerable because of its circular symmetry. In [59], Freeman and Adelson have shown that the first order x-derivative (G_1^θ) of a Gaussian filter oriented at an arbitrary orientation θ can be expressed as a linear combination of ($G_1^{0^o}$) and ($G_1^{90^o}$) in the following manner:

$$G_1^\theta = cos(\theta)G_1^{0^o} + sin(\theta)G_1^{90^o} \tag{2.3}$$

In the above equation, $G_1^{0^o}$ and $G_1^{90^o}$ are the basis filters and the terms $cos(\theta)$ and $sin(\theta)$ are the interpolation functions. Thus an image filtered at any orientation can be expressed as a linear combination of the image convolved with the basis filters

(convolution operation being linear).

Then, we can write:

$$R_1^{0^o} = G_1^{0^o} * I \tag{2.4}$$

$$R_1^{90^o} = G_1^{90^o} * I \tag{2.5}$$

$$R_1^{\theta} = cos(\theta) R_1^{0^o} + sin(\theta) R_1^{90^o} \tag{2.6}$$

In the above equation, $R_1^{\theta^o}$ represents the image I filtered using the basis filter at an arbitrary orientation θ and $*$ denotes the convolution operation.

2.3.3 Image saliency models

Saliency model aims to capture visually attentive regions in an image [62]. A single approach is incapable of detecting all the salient regions accurately for all images [63]. However, a synergistic combination of some of the highly performing saliency methods can lead to a highly informative and accurate saliency map. The proposed model uses a linear weighted combination of three saliency maps obtained using Graph-based visual saliency (GBVS), Laplacian saliency (LS) and Spectral residual saliency (SRA). Now, we briefly discuss about the individual saliency models.

GBVS uses topological structure of the graphs to compute saliency values and employs Markovian approach in the process [64]. This model comprises of three stages: extraction of the important feature vectors from the scene, construction of an "activation map(s)", and, combination of these maps for obtaining a single saliency map. The construction of activation maps is based on a linear filtering method [65]. Markovian approach is employed to construct the activation map. An elegant dissimilarity measure is adopted to identify locations having high variations. The dissimilarity of

$M(i,j)$ and $M(p,q)$ can be mathematically expressed as:

$$d(i,j) \parallel (p,q) \triangleq |\log\left(\frac{M(i,j)}{M(p,q)}\right)| \tag{2.7}$$

The normalized maps are summed over each feature channel to obtain the master saliency map.

LS uses Laplacian filter to capture the high frequency regions (salient regions) in the image [66]. It computes the saliency map S, of image I, as the local average of the absolute value of high-pass image H_I obtained using convolution of the input image I with 3×3 Laplacian mask using:

$$S = |H_I| * g_{r_g,\sigma_g} \tag{2.8}$$

Where, $H_I = I * L$ and g is a Gaussian low-pass filter of size $(2r_g + 1).(2r_g + 1)$; r_g and σ_g have been set to 5.

SRA employs the power of log spectrum for saliency detection [67]. The log spectrum representation of an image, $L(f)$ can be expressed in terms of the amplitude of the Fourier spectrum, $A(f)$ of that image, (f denotes frequency) using:

$$L(f) = \log(A(f)) \tag{2.9}$$

The algorithm aims at reducing the redundant information in the image. It focuses on the statistical singularities in the spectrum. The statistical singularities, also defined as the spectral residual of an image, can be obtained using:

$$R(f) = L(f) - A(f) \tag{2.10}$$

In (12), $R(f)$ represents the spectral residual of the image. The spectral residual is further processed to construct the saliency map in spatial domain by using inverse

Fourier transform.

2.3.4 Guided filtering

Very recently guided filtering technique [68] has gained prominence in computer vision applications like edge preserving smoothing, dehazing, feathering and image matting. *Guided filter* proposed by He et al. [68, 69] is an edge-preserving smoothing filter, derived from a local linear model between guidance I and output q. The guided filter computes the output by taking into account the content of the guidance image which can be the filtering input itself [70, 71]. Assuming q as a linear transform of I in window ω_k centered at pixel k, q can be expressed for a pixel i as below.

$$q_i = a_k I_i + b_k \quad \forall \ i \in \omega_k \tag{2.11}$$

Here, (a_k, b_k) are linear coefficients assumed to be constant in the window ω_k. As $\bigtriangledown(q) = a.\bigtriangledown(I)$, the guided filter preserves edges and can be used in applications like image matting, dehazing and feathering [68]. The linear coefficients are determined using constraints between filtering input p and q given as:

$$q_i = p_i - n_i \tag{2.12}$$

Where, n_i refers to unwanted noise components. The linear ridge regression model is used to optimize the cost function by minimizing the difference between p and q. The coefficients are given by:

$$a_k = \frac{\frac{1}{|\omega|}\sum_{i\in\omega_k} I_i p_i - \mu_k \bar{p_k}}{\sigma_k^2 + \in} \tag{2.13}$$

$$b_k = \bar{p_k} - a_k \mu_k \tag{2.14}$$

Here, μ_k represents the mean of I in window ω_k and σ_k^2 represents the variance of the image in the same window. $|\omega|$ is the number of pixels in the window ω_k and $\bar{p_k}$ is the mean of p computed in that window. As a pixel i can be involved in overlapping windows, the final output should be obtained by taking the average of all possible values of q.

$$q_i = \bar{a_i} I_i + \bar{b_i} \tag{2.15}$$

Where, $\bar{a_i}$ and $\bar{b_i}$ are the average value of a_k and b_k of all windows overlapping i.

2.4 Proposed method

2.4.1 Steerable local frequency based solution

Oppenheim et al. [54] demonstrated the importance of phase in images through a series of experiments. The standard focus measures are mainly based on intensity. In this work, we explore the potential of local phase information of the pixels in the source images for determining the focus measure. Features in an image can be oriented at any angle θ ($0^o \leq \theta \leq 180^o$) [72]. For each pixel, corresponding responses from the filter at different orientations need to be compared to get the maximum response. Local frequency map obtained from the analytic image does not include any knowledge of orientation. To capture orientation, we introduce the concept of steerable local frequency map. Oriented analytic image is used to build the steerable local frequency map. Hilbert transform, realized through the quadrature pair of filters (G_4, H_4), is used first to obtain the analytic image. Fourth order derivative of Gaussian (G_4) offers higher resolution analysis as it has narrow frequency tuning. In [59], the approximation to the Hilbert transform of G_4, denoted by H_4, is obtained using the least squares fit of product of a 5^{th} order polynomial with six basis functions and a radially symmetric Gaussian function. To obtain the oriented analytic image, we therefore require a steerable Hilbert kernel. However, since the Hilbert Transform

itself cannot be made steerable in its present form, we apply the concept of steerable quadrature pair of filters (G_4^θ, H_4^θ). The analytical expression of G_4^θ is given by:

$$G_4^\theta = [K_a(\theta)G_4a + K_b(\theta)G_4b + K_c(\theta)G_4c + K_d(\theta)G_4d + K_e(\theta)G_4e] \tag{2.16}$$

where, G_4a, G_4b, G_4c, G_4d and G_4e constitute the basis set functions and $K_a(\theta)$, $K_b(\theta)$, $K_c(\theta)$, $K_d(\theta)$ and $K_e(\theta)$ are the interpolation functions. Similarly, the analytic expression for and H_4^θ is given by:

$$H_4^\theta = [K_a(\theta)H_4a + K_b(\theta)H_4b + K_c(\theta)H_4c + K_d(\theta)H_4d + K_e(\theta)H_4e + K_f(\theta)H_4f] \tag{2.17}$$

here, H_4a, H_4b, H_4c, H_4d, H_4e and H_4f are basis set functions and $K_a(\theta)$, $K_b(\theta)$, $K_c(\theta)$, $K_d(\theta)$, $K_e(\theta)$ and $K_f(\theta)$ are the interpolation functions. The equations for basis and interpolation functions are given in Appendix A. This steerable quadrature pair G_4^θ and H_4^θ is used to filter the original image $I(x, y)$ to obtain the oriented analytic image $I_{A,\theta}(x, y)$ at an arbitrary orientation θ. So, we can write:

$$I_{A,\theta}(x, y) = I_{G4,\theta}(x, y) - jI_{H4,\theta}(x, y) \tag{2.18}$$

$$I_{G4,\theta}(x, y) = I(x, y) * G_4^\theta \tag{2.19}$$

$$I_{H4,\theta}(x, y) = I(x, y) * H_4^\theta \tag{2.20}$$

$I_{G4,\theta}(x, y)$ and $I_{H4,\theta}(x, y)$ together constitute the steerable quadrature filtered response of the original image $I(x, y)$. The steerable local phase $\phi_\theta(x, y)$ of the Gaussian

filtered image can now be obtained using.

$$\phi_\theta(x,y) = abs\left[arc\tan\left\{\frac{I_{H4,\theta}(x,y)}{I_{G4,\theta}(x,y)}\right\}\right] \tag{2.21}$$

To get rid of background noise or distortions, the mean of the steerable local phase map is subtracted from the phase value at each of the pixels to construct the modified phase map $\phi_\theta^{'}(x,y)$ [72].

$$\phi_\theta^{'}(x,y) = \phi_\theta(x,y) - \overline{\phi_\theta} \tag{2.22}$$

In the above equation, $\overline{\phi_\theta}$ is mean of the steerable local phase map. Steerable local frequency map is obtained using the gradient of the modified local phase in the following manner:

$$Freq_\theta(x,y) = \sqrt{\left[\frac{\partial(\phi_\theta^{'}(x,y))}{\partial x}\right]^2 + \left[\frac{\partial(\phi_\theta^{'}(x,y))}{\partial y}\right]^2} \tag{2.23}$$

where,

$$(\partial\phi_\theta^{'}(x,y)/) = \partial\phi_\theta^{'}(x+1,y) - \partial\phi_\theta^{'}(x,y) \tag{2.24}$$

$$(\partial\phi_\theta^{'}(x,y)/\partial y) = \partial\phi_\theta^{'}(x+1,y) - \partial\phi_\theta^{'}(x,y) \tag{2.25}$$

The local frequency maps obtained at different orientations are further max-pooled to obtain the resultant steerable local frequency map, $Freq_{\theta_{max}}(x,y)$.

$$Freq_{\theta_{max}}(x,y) = max[Freq_{\theta_1}(x,y), Freq_{\theta_2}(x,y), Freq_{\theta_3}(x,y),, Freq_{\theta_{13}}(x,y)] \tag{2.26}$$

In the above equation, $\theta_1, \theta_2, \ldots, \theta_{13}$ denote 13 orientations covering the entire range [0^o, 180^o] in steps of 15^o. Number of orientations is chosen experimentally and this

is described later, see section 2.5.3.2. We then choose the best suitable threshold T experimentally, for thresholding max-pooled local frequency map, $Freq_{\theta_{max}}$, to compute the number of interest points in source image. Selection of the best performing threshold (T) is also discussed later, see section 2.5.3.1. Number of interest points n detected in the source image is given by:

$$n = \sum_{x} \sum_{y} [Freq_{\theta_{max}}(x, y)) \geq T] \tag{2.27}$$

Thus, the proposed focus measure can be normalized in [0, 1] using [23]:

$$FM_{proposed} = \left[\frac{n - n_{min}}{n_{max} - n_{min}} \right] \tag{2.28}$$

Here, n_{max} is the maximum number of interest points and n_{min} is the minimum number of interest points detected among all the source images in a set I

We now present Algorithm 1 where various steps to obtain the proposed focus measure are shown.

Algorithm 1 Computation of Focus Measure (FM)

Input: I, An image from Visible(VIS), Near-Infrared(NIR) or Thermal(TH) image set.

Output: $FM_{proposed}$

Initialization: $n = 0$, $[p, q]$=$size(I)$

for θ = $0^o : 15 : 180^o$ **do**
  Obtain 7×7 G_4^θ and H_4^θ at θ
  Obtain $I_{G4,\theta}$ and $I_{H4,\theta}$
  Obtain oriented analytic image, $I_{A,\theta}(x, y)$
  Obtain steerable local phase map, $\phi_\theta(x, y)$
  Compute, $\phi'_\theta(x, y) = \phi_\theta(x, y) - \overline{\phi_\theta}$
  Obtain Steerable local frequency, $Freq_\theta(x, y)$
end for
Obtain $Freq_{\theta_{max}}(x, y)$
Obtain threshold, T (experimentally)
for i=$1 : p$ **do**
  for j=$1 : q$ **do**
    if $Freq_{\theta_{max}}(i, j) \geq T$ **then**
      $n = n + 1$
    end if
  end for
end for
Obtain $FM_{proposed}$

The proposed multifocus image fusion (SLF method) begins with the assumption that the source images to be fused are pre-registered. The resultant fused image F contains pixels from the source image having highest max-pooled local frequency value for that pixel. Further, a 3×3 majority filter is applied for consistency verification [56]. This step ensures that a pixel in the fused image is not allowed to come from a source image if majority of its neighbors in the two images (fused and the source) are different. This measure plays an important role in characterizing the performance of image fusion algorithms.

We, next, present Algorithm 2 where various steps for proposed SLF based multifocus image fusion in VIS and NIR spectra are given.

Algorithm 2 Multifocus Image Fusion (SLF method): For VIS and NIR spectrum

Input: I_N, Multifocus source image set of N images

Output: F, All-in-focus image

Initialization: $[p, q]$=$size(I_N)$

1: **for** k = 1 : N **do**

2: Obtain $Freq_{\theta_{max},k}$ for k^{th} image, using Algorithm 1 (Step 1-9)

3: **end for**

4: **for** i=1 : p **do**

5: **for** j=1 : q **do**

6: Q=$arg\max_{k=1:N}(Freq_{\theta_{max},k}(i,j))$

7: $F(i,j) = I_Q(i,j)$

8: **end for**

9: **end for**

10: Perform consistency verification using 3x3 majority filter (Optional)

For thermal multifocus image fusion, we use proposed focus measure as activity level and formulate weighted average based fusion rule, as in recently reported EOL based method for thermal spectrum [52]. This is just for fair comparisons with only first of its kind multifocus thermal image fusion method [52]. The variant of the algorithm is given in Algorithm 3.

Algorithm 3 Multifocus Image Fusion (SLF method): For TH spectrum

Input: I_N, Multifocus source image set of N images

Output: F, All-in-focus image

Initialization: $[p, q]$=$size(I_N)$

1: **for** k = 1 : N **do**

2: Obtain $Freq_{\theta_{max},k}$ for k^{th} image, using Algorithm 1 (Step 1-9)

3: **end for**

4: **for** i=1 : p **do**

5: **for** j=1 : q **do**

6: $\omega_k(i,j) = \frac{Freq_{\theta_{max},k}(i,j)}{\sum_{k=1}^{N} Freq_{\theta_{max},k}(i,j)}$

7: $F(i,j) = \sum_{k=1}^{N} \omega_k(i,j) * I_k(i,j)$

8: **end for**

9: **end for**

2.4.2 Guided SLF and improved saliency model based solution (GSLF-IS)

We further enhance the performance of our multispectral multifocus image fusion method by using novel combination of improved saliency model and guided filtering. We propose judiciously application of the guided edge preserving filtering in two phases. In the first phase, the source images to be fused are enhanced using guided filtering keeping the source images same as the guidance images. Steerable local frequency map (SLF) capturing phase/frequency information at different orientations is suggested in section 2.4.1. In the second phase, the steerable local frequency maps of the enhanced source images are further refined using guided filtering. In this case, the enhanced source images are used as the guidance images. We further develop an improved model of saliency based on Graph-based visual saliency (GBVS) [64], Spectral residual saliency (SRA) [67] and Laplacian saliency (LS) [66]. This improved saliency map is combined with the guided steerable local frequency (GSLF) map to generate good fusion results across all spectra.

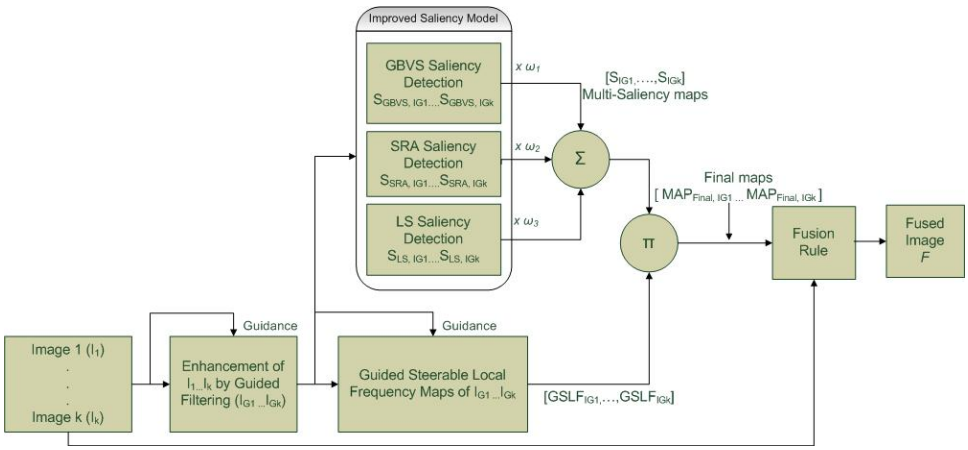


FIGURE 2.2: Schematic diagram of GSLF and improved saliency based solution

Fig. 2.2 shows the block diagram of our solution pipeline. The proposed method starts with guided filtering of the source images which are further processed to yield improved saliency maps and steerable local frequency (SLF) maps. SLF map is enhanced with guided filtering. The composite saliency map is combined with guided SLF (GSLF) map to yield improved fusion results. The integration aims at incorporating the combined effect of intensity, phase and orientation. The major components

governing the above process are discussed in the following subsections.

2.4.2.1 Proposed image saliency model

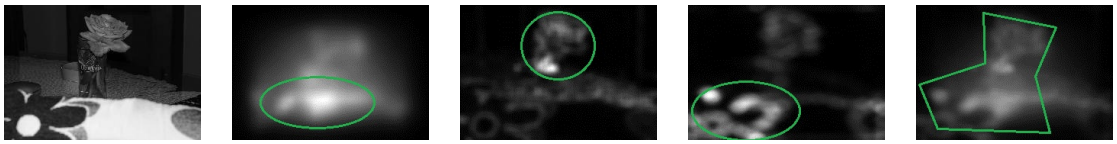
Saliency model aims to capture visually attentive regions in an image [62]. A single approach is incapable of detecting all the salient regions accurately for all images [63]. However, a synergistic combination of some of the highly performing saliency methods can lead to a highly informative and accurate saliency map. The proposed model uses a linear weighted combination of three saliency maps obtained using Graph-based visual saliency (GBVS), Laplacian saliency (LS) and Spectral residual saliency (SRA).

The integration realizes the full potential of the three individual schemes to obtain an improved saliency model. Note that GBVS is a robust and computationally efficient scheme owing to the use of graphs [64]. LS uses the Laplacian operator which uses second order derivative to determine the edges in an image. SRA model offers a general solution for salient region detection [67]. We construct the final saliency map of the input image I, S_I by taking weighted combination of the three saliency maps, $S_{GBVS,I}$, $S_{SRA,I}$ and $S_{LS,I}$ in the following manner:

$$S_I = \sum_{m=1}^{3} \omega_m N(S_{I,m}) \tag{2.29}$$

In equation (2.29), $\sum_{m=1}^{3} \omega_m = 1$, $N(S_{I,m})$ denotes the m^{th} normalized saliency map and ω_m is its corresponding weight [63]. In the proposed method we use weights $\omega_{GBVS} = 0.5$, $\omega_{LS} = 0.3$ and $\omega_{SRA} = 0.2$ and construct the final saliency map. These weights are determined experimentally and kept constant throughout.

An example of improved saliency detection using the proposed model is illustrated in Fig. 2.3. The test image is shown in Fig 2.3(a). Green polygons in each of the figures (b), (c) and (d) indicate the dominant salient regions. These images clearly show that the obtained saliency maps are complementary in nature. So, no single method can



(a) (b) (c) (d) (e)

FIGURE 2.3: Example of improved Saliency map : (a) Test image (b) GBVS saliency map (c) LS saliency map (d) SRA saliency map (e) Proposed method. Green polygons indicate dominant salient regions.

detect all the dominant salient regions. As a marked improvement, figure (e) shows that the map obtained using the proposed saliency model can successfully capture all the perceptually salient regions in the test image that also includes edges.

2.4.2.2 Guided steerable local frequency

As stated earlier, we employ guided filter in two stages. In the first stage, the input images are enhanced. Our objective is to obtain a high quality fused image which is definitely dependent on the quality of input. So, to obtain improved inputs, we apply the linear model of guided filter [68]. The guidance image in this case is the input image itself (I). The filtered output I_G at pixel i gives the guided input as shown below:

$$I_{G,i} = a_k I_i + b_k \quad \forall \ i \in \omega_k \tag{2.30}$$

$$I_{G,i} = I_i - n_i \tag{2.31}$$

The guided input obtained as a result is used for the construction of SLF and saliency maps. The second stage attempts at modifying the steerable local frequency maps using guided filter. The objective here is to improve the maps which will enable efficient representation of features with increased accuracy. The guided filter accepts the SLF map as the original filtering input and treats the guided input obtained from first phase as the guidance image. The filtered output then yields finely tuned feature map. The mathematical representation describing the process for a pixel i is given

below:

$$GSLF_{I_G,i} = a_k I_{G,i} + b_k \quad \forall \ i \in \omega_k \tag{2.32}$$

$$GSLF_{I_G,i} = SLF_{I_G,i} - n_i \tag{2.33}$$

Here, SLF_{I_G} is the original steerable local frequency map of the guided input image I_G and $GSLF_{I_G}$ is its improved guided version. The same guided input image I_G acts as the guidance image. The reason behind improvement in SLF maps using guided filter is explained next. Steerable local frequency which primarily uses Hilbert transform is linearly related to guidance image (Hilbert transformer being a linear time-invariant filter). The guidance image for our work is the guided input which preserves the gradients. From equation 2.32 it is apparent that the filtering output is basically a scaled version of the guidance image displaced by an offset. The local linear model of the guided filter supports structure-transfer filtering due to its patch-based model [68]. This unique property enables the transfer of fine structures present in the guided input to $GSLF_{I_G}$, even if the original filtering input is smooth in some regions. Thus, an enhanced steerable local frequency map containing sharp features is obtained.

2.4.2.3 Fusion

Let S_{I_G} be the saliency map of the improved input. For each source image, we combine $GSLF_{I_G}$ and S_{I_G} by taking their product. The result yields the final map, MAP_{Final}, for each of the multifocus source image. So, we can write:

$$MAP_{Final,i} = (GSLF_{I_G,i}) \times (S_{I_G,i}) \tag{2.34}$$

For the VIS and the NIR spectrum, the fused image F contains pixels belonging to the source image possessing highest corresponding value in MAP_{Final} for that particular

pixel i.

$$Q = \arg\max_{k}(MAP_{Final,i,k}) \tag{2.35}$$

Where, $k \in [1, N]$. Here N is the total number of source images to be fused and Q is the index of source image having maximum value of MAP_{Final} for the pixel i. So, the fused image F is obtained by choosing each pixel i from the most suitable source image with index Q. So, we can write:

$$F = \bigcup_{i} F_i = \bigcup_{i} I_{Q,i} \tag{2.36}$$

Further, a 3×3 majority filter is applied for consistency verification [56] to ensure that a pixel in the fused image does not come from a source image different from that of its majority of neighbors. In case of TH spectrum, we use pixel-level weighted averaging rule to obtain final fused result F as in [73], to avail better comparison.

Please note that the methods with which we have compared our work in the NIR and the TH spectra are obtained from applying different fusion strategies. So, in order to have proper comparisons with these methods from different spectra, we have to fuse our NIR and TH images accordingly.

2.5 Experimental results

In this section, we first mention the datasets used for various experimentation along with the performance evaluation measures/metrics. We next discuss how certain parameters (thresholds) are chosen experimentally. We then show the comparative performance analysis of the proposed SLF based focus measure in different spectra like VIS, NIR and TH. Next, we demonstrate the improvements in fusion results using our focus measure. Further, we also demonstrate the enhancement in the performance by introducing the improved saliency model and the use of guided filtering to enhance the source images to be fused and steerable local frequency maps.

2.5.1 Evaluation Dataset

For the evaluation of focus measure performance we use three multispectral datasets, one each from the visual (VIS), near-infrared (NIR) and thermal (TH) spectrum [23]. Each dataset in turn consists of seven sets of images. Some sample image sets are shown in Fig. 2.4. For the evaluation of the multifocus image fusion in the visual spectrum, we use the same image sets as in [37] (see Fig. 2.5). In the near-infrared, only one image set was found to be suitable from the available NIR dataset [23] (see Fig. 2.6). We also use a multimodal medical image set to evaluate proposed fusion method. This consists of CT and MRI modality image set of human brain and the images are shown in Fig. 2.7. For the thermal spectrum, we experiment with the reduced set of multifocus thermal image datasets developed by Benes et al. [52]. The original thermal image database consists of five multifocus image sets with 96 images in each set. All the sets contain a scene image with two objects but with different backgrounds, varying temperatures and different object distances. A reduced set of 10 images for each dataset is derived from the original pool of 96 images using EOL based activity level measurement [52]. The reduced image sets for the mobile-interface and the two bulbs are shown in Fig. 2.8 and Fig. 2.9.

2.5.2 Performance measures

The performance evaluation measures are now briefly discussed below. The focus measure is evaluated based on different criteria such as monotonicity, magnitude of slope and smoothness. For this we employ the Q (Quality factor) and P (Peak of focus curve) performance metrics [23].

1. Q (Quality factor): The quality factor is computed from the focus curve. The focus curve is the plot between image index (N) and focus measure ($FM_{proposed}$).

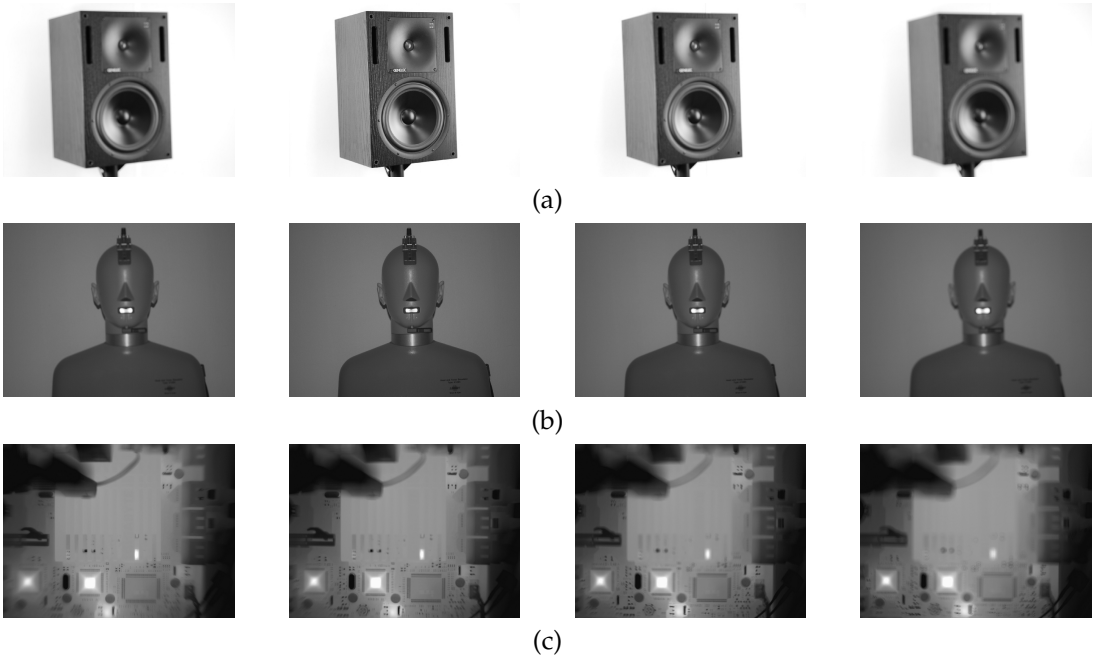


FIGURE 2.4: Sample images from each multispectral dataset used for focus measure evaluation: (a) 'Loudspeaker' (VIS), (b) 'Head' (NIR), (c) 'Circuit' (TH).

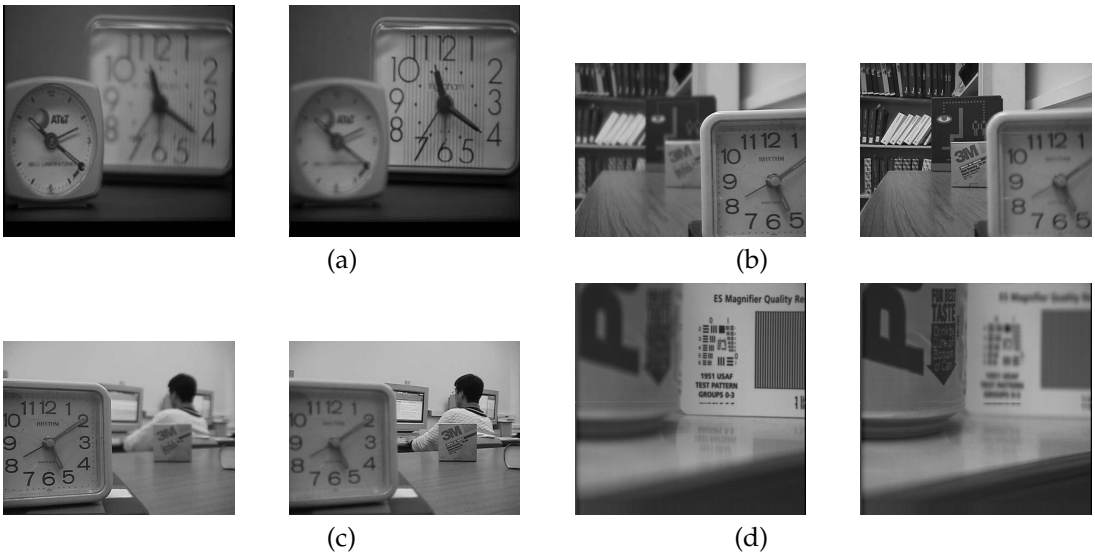


FIGURE 2.5: Visual multifocus image datasets used for evaluation of the proposed multifocus image fusion methods: (a) 'Clock', (b) 'Desk', (c) 'Lab', (d) 'Pepsi'

The formula for Q is given below:

$$Q = \left[\frac{1}{N_{max} - N_{min} + 1} \right] \tag{2.37}$$

$$C_s[N] \geq 0.7079, \quad For \;\; N_{min}, \ldots, N, \ldots, N_{max} \tag{2.38}$$

FIGURE 2.6: Near-infrared (NIR) multifocus image dataset used for evaluation of proposed multifocus image fusion methods: 'keyboard'

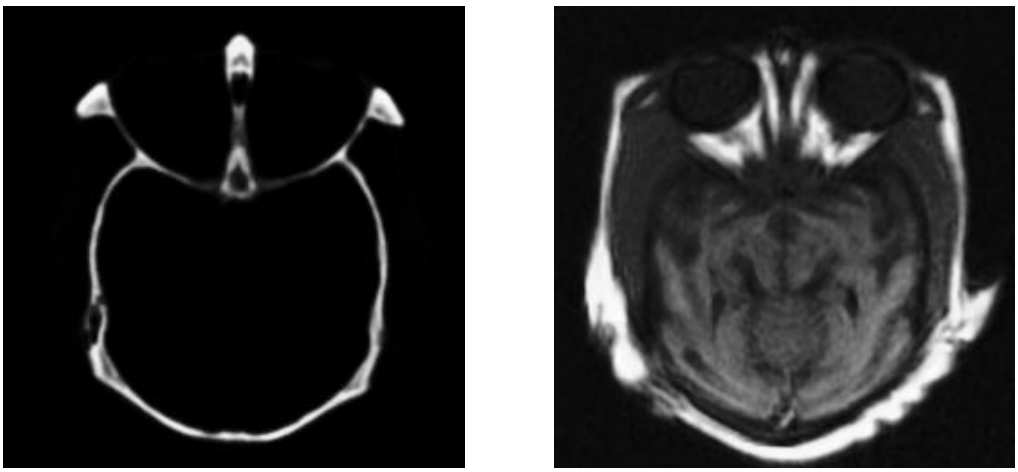


FIGURE 2.7: Medical image dataset of Brain for the evaluation of proposed image fusion method. (a) CT (b) MRI.

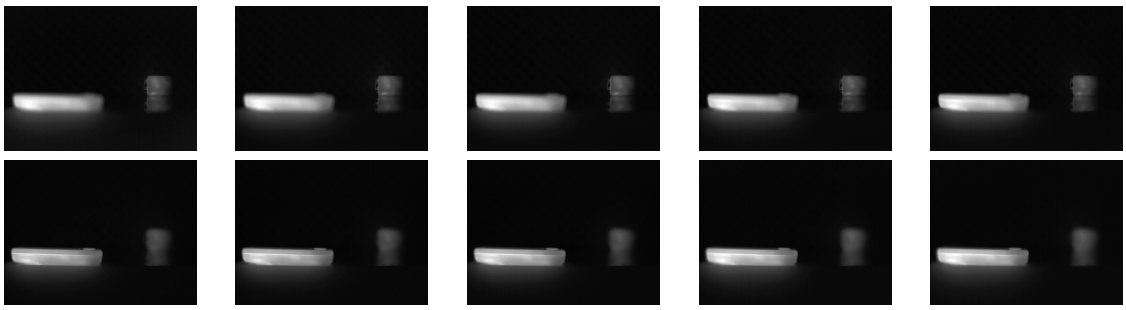


FIGURE 2.8: Reduced thermal multifocus image dataset used for the evaluation of proposed multifocus image fusion methods: Set 1 (Mobile phone and RS 232 interface).

$C_s[N]$ in equation 2.38 is the focus curve normalized in the range $[0, 1]$. Number of focus curve samples higher than 0.7079 are used to measure the Q factor. A narrow peak in the focus curve with a high Q-factor is favorable.

2. P (Peak of focus curve): P represents the image having highest focus evaluated from the focus curve, $C_s[N]$.

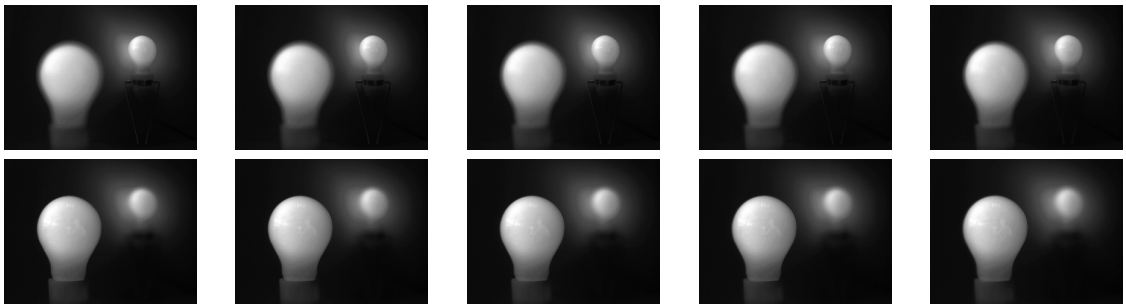


FIGURE 2.9: Reduced thermal multifocus image dataset used for the evaluation of proposed multifocus image fusion methods: Set 2 (Two bulbs).

For the evaluation of multifocus image fusion in the visual and near-infrared spectrum, we use MI (Mutual Information), $Q^{AB/f}$ and Q_0. They are described below:

1. MI (Mutual Information) [74]: MI measures the statistical dependence between two random variables and the amount of information that one variable contains about the others. Here, the MI between source images A and B and fused image F is given by:

$$MI = I_{AF} + I_{BF} \tag{2.39}$$

In equation 2.39 I_{AF} is the mutual information between the source image A and the fused image F whereas I_{BF} is the mutual information between the source image B and the fused image F. A high value of MI indicates better result.

2. $Q^{AB/f}$ [75]: This metric reflects the quality of visual information obtained from the fusion of input images. $Q^{AB/f}$ can be defined as:

$$Q^{AB/f} = \frac{\sum_{n=1}^{N}\sum_{m=1}^{M}(Q^{AF}(n,m)w^{A}(n,m) + Q^{BF}(n,m)w^{B}(n,m))}{\sum_{n=1}^{N}\sum_{m=1}^{M}(w^{A}(n,m) + w^{B}(n,m))} \tag{2.40}$$

In equation 2.40, A and B denotes the source images and f denotes the final fused image. Q_{AF} and Q_{BF} represents amount of edge information preserved in F from image A and that from image B respectively. w^A and w^B are weights derived by convolving Sobel operator with images A and B [75]. $Q^{AB/f}$ varies

in the range [0, 1] where a value of 1 corresponds to the best performance.

3. Q_0 [76]: This metric is designed by modeling any image distortion as a combination of three factors, namely, loss of correlation, luminance distortion, and contrast distortion. The value of Q_0 between source images A, B and fused image F is expressed as:

$$Q_0(A, B, F) = \left[\frac{Q_0(A, F) + Q_0(B, F)}{2}\right] \tag{2.41}$$

where, $Q_0(A, F)$ is defined as :

$$Q_0(A, F) = \left[\frac{\sigma_{af}}{\sigma_a \sigma_f} \cdot \frac{2af}{(a)^2 + (f)^2} \cdot \frac{2\sigma_a \sigma_f}{(\sigma_a^2 + \sigma_f^2)}\right] \tag{2.42}$$

Here, σ_a and σ_f are standard deviations of input image A and fused image F; σ_{af} denotes the covariance between A and F. The dynamic range of $Q_0(A, B, F)$ is $[-1, 1]$ with best possible value as 1.

Three metrics $RMSE$ (Root Mean Square Error), MAE (Mean Absolute Error) and CC (Cross Correlation) as in [52], are employed to evaluate the performance of the proposed fusion method in thermal spectrum and are described below.

4. $RMSE$: The Root Means Square Error between the fused image F and reference ground truth image R is given by.

$$RMSE = \sqrt{\frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} |R(i, j) - F(i, j)|^2} \tag{2.43}$$

Here, NM is size of the image. Lower the value of $RMSE$, better is the performance of fusion.

5. MAE: The Mean Absolute Error between the fused image F and reference ground truth image R is given by.

$$MAE = \sqrt{\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}|R(i,j) - F(i,j)|} \tag{2.44}$$

   Lower the value of MAE, better is the performance of fusion.

6. CC: The Cross Correlation between the fused image F and reference ground truth image R can be expressed as.

$$CC = \frac{2\sum_{i=1}^{N}\sum_{j=1}^{M}R(i,j)F(i,j)}{\sum_{i=1}^{N}\sum_{j=1}^{M}R(i,j)^2 + \sum_{i=1}^{N}\sum_{j=1}^{M}F(i,j)^2} \tag{2.45}$$

   The dynamic range of CC is [0, 1] with best possible value as 1.

2.5.3 Selection of Threshold and Number of orientations to obtain image level focus measure

We perform experiments to judiciously select the number of orientations (O) and the threshold (T) in the proposed method.

2.5.3.1 Selection of Threshold

We have experimentally obtained the best performing value of threshold parameter T. For the range of T, we used [min, max] of max-pooled local frequency map. T is set from the above range based on the performance of the focus curves ($C_s[N]$) [55] in terms of Accuracy, Width at 50% maximum and Number of local maxima. Some sample focus curves obtained ($C_s[N]$) for five different threshold values ($T1$, $T2$, $T3$, $T4$, $T5$) in this range are shown in Fig. 2.10, 2.11 and 2.12. The focus measure curves reveal an interesting trend. The curves are almost comparable with varying values of threshold indicating the robustness of the proposed focus measure. However, in

terms of Accuracy and Width at 50% maximum, the experimentally selected value of $T3$ (=0.0607) emerges as an optimal choice for the threshold.

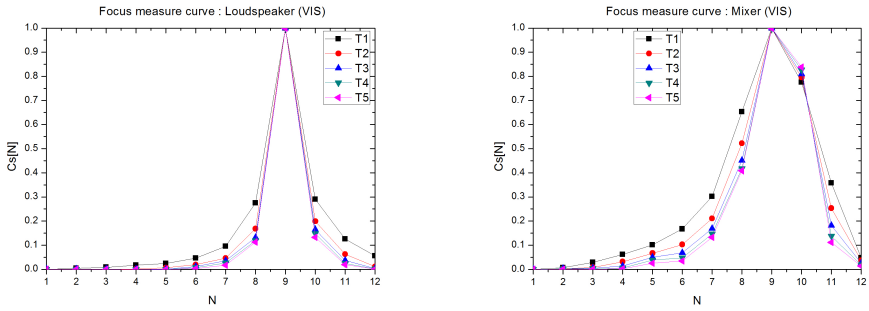


FIGURE 2.10: Specimen focus curves for 'Loudspeaker' and 'Mixer' image sets in VIS spectrum for different threshold values.

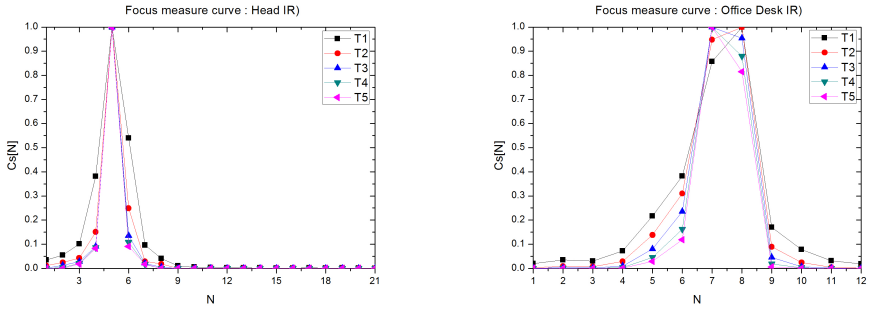


FIGURE 2.11: Specimen focus curves for 'Head' and 'Office desk' image sets in NIR spectrum for different threshold values.

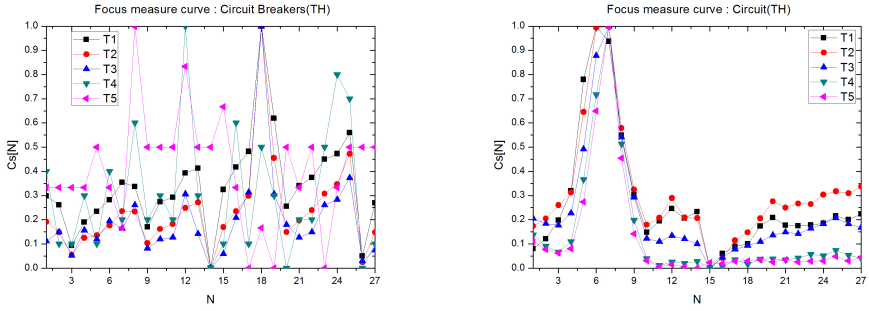


FIGURE 2.12: Specimen focus curves for 'Circuit breakers' and 'Circuit' image sets in TH spectrum for different threshold values.

2.5.3.2 Selection of Number of orientations

Features in an image can be oriented at any angle θ within the range 0^o to 180^o [72]. Selection of number of intervals for orientations influences the detection of oriented features in the input image. Note that less number of intervals may fail to capture the

finer oriented features present in the image. On other hand, use of large number of intervals for orientations can be unreliable (sensitive to noise) in addition to increasing the computational overhead. So, as a trade-off, five intermediate choices $O1$ (7 orientations in steps of 30^0), $O2$ (10 orientations in steps of 20^0), $O3$ (13 orientations in steps of 15^0), $O4$ (16 orientations in steps of 12^0), $O5$ (19 orientations in steps of 10^0) are used with a fixed threshold (T). The focus measure curves ($C_s[N]$) for different orientations for some sample image sets from visible (VIS), near infra-red (NIR) and thermal (TH) spectra are shown in Fig. 2.13, 2.14 and 2.15.

These focus measure curves are evaluated in terms of Accuracy, Width at 50% maximum and Number of local maxima of the focus curves obtained ($C_s[N]$) [55]. In the visible (VIS) spectrum the performance is uniform with respect to Accuracy but in terms of Width at 50% maximum $O3$ yields better performance. As can be seen from the focus curves in the NIR spectrum, $O3$ performs better. For example, in case of 'Office Desk' image set $O3$ produces peak at image index 7 which is nearest to the index of image having highest focus from subjective assessment test given in [48]. In the TH spectrum there are a number of local maxima for each number of orientations due to limited resolution. But based on Width at 50% maximum and Accuracy it is clear that $O3$ performs better. So, we have used $O3$ (13 orientations in step of 150) for our work.

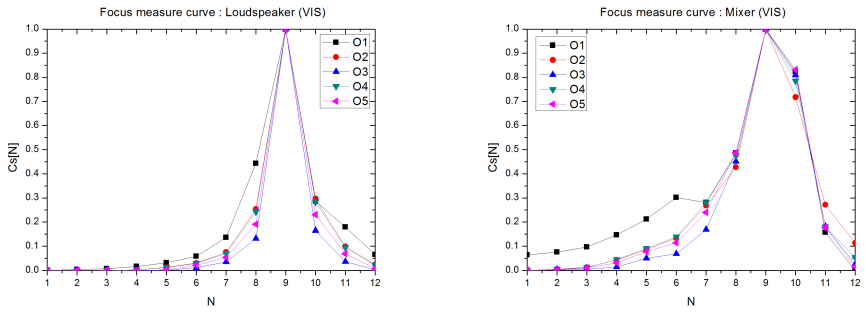


FIGURE 2.13: Specimen focus curves for 'Loudspeaker' and 'Mixer' image sets in VIS spectrum for different number of orientations.

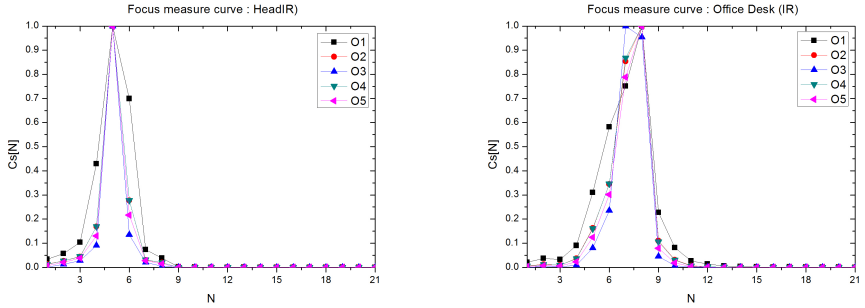


FIGURE 2.14: Specimen focus curves for 'Head' and 'Office desk' image sets in NIR spectrum for different number of orientations.

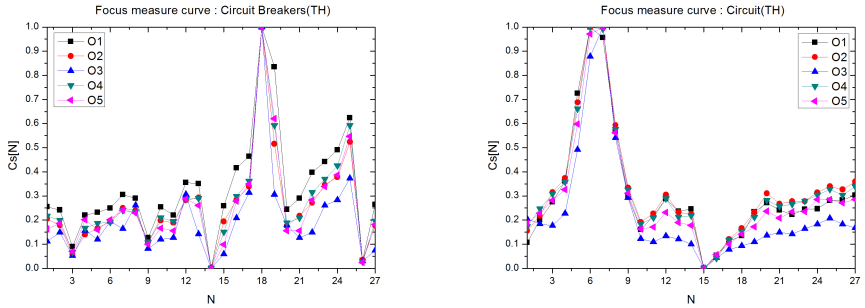


FIGURE 2.15: Specimen focus curves for 'Circuit breakers' and 'Circuit' image sets in TH spectrum for different number of orientations.

2.5.4 Performance comparison for SLF based multispectral focus measure

We compare our proposed focus measure with other such measures from two categories, namely, i) the standard intensity driven focus measures and ii) the recently developed IPD based focus measures. The first category includes Energy of Laplacian (EOL), Sum Modified Laplacian (SML) and Spatial Frequency (SF). From the second category we compare with Fast Hessian (FH), Harris-Laplace (HL) and Features from Accelerated Segment Test (FAST) [23]. The standard focus measures are known to yield good results. EOL and its modified adaptation SML are high performing derivative based focus measures. On the other hand IPD based focus measures perform relatively well when applied to multispectral images.

We now show focus curves for the proposed focus measure ($C_s[N]$) for datasets from the different spectra. Please see Fig. 2.16 for the focus curves of the 'Loudspeaker' and the 'Mixer' datasets in the visual spectrum, Fig. 2.17 for the focus

curves of the 'Head' and the 'Office desk' datasets in the near-infrared spectrum, and, Fig. 2.18 for the focus curves of the 'Circuit breakers' and the 'Circuit' datasets in the thermal spectrum. A good focus measure possesses the characteristics of unimodality, monotonicity and is sensitive to defocus [23]. Our method exhibits all the desirable characteristics warranted of a good focus measure. Focus curves evaluated for the proposed focus measure reach a global maximum and decrease monotonically as the defocus increases on either side. However, a few false maxima and minima are observed in the focus curves of thermal images because of poor resolution due to limited focal length. It is reported in [23] that the interest point detectors perform dismally in the visual spectrum as compared to the standard focus measures. However, their performance is improved substantially in other spectra (thermal, near-infrared). Our interest point based focus measure shows decent performance across all the spectra. Comparative results for the visual spectra are shown in Table 2.1. EOL, SML and SF perform well for most of the datasets. Compared to other interest point detectors, the proposed IPD based focus measure outperforms FAST, FH and HL in most of the cases. The proposed method is comparable with subjective analysis in terms of the P metric as reported in [23]. Results for the near-infrared spectrum in Table 2.2 show significant improvements in the performance over other interest point detector based focus measures as well as standard focus measures. The Keyboard dataset belonging to near-infrared spectrum reveals that the proposed focus measure yields a Q value of 0.5 whereas the reported Q values of FH, FAST and HL are 0.25, 0.10, and, 0.1250 respectively. Our Q value is comparable to that of EOL and is better than SML and SF with reported values as 0.0830 and 0.3333. We outperform SML, SF in most of the cases. We have outdone FH in all the cases and perform much better compared to FAST and HL. Comparative analysis reveals that the performance of the proposed focus measure is best for the thermal images as shown in Table 2.3. Our detector performs better than SML, SF, FAST and HL and is comparable to EOL and FH. In the 'Circuit breakers' dataset, the best performing interest point detector based focus measure is FH and that from the standard focus measures is EOL,

each having a Q value of 0.5. The proposed focus measure having Q value of 1.0 easily surpasses both. The superior performance of our focus measure as compared to the standard interest point detector based focus measures is due to use of phase information at various orientations.

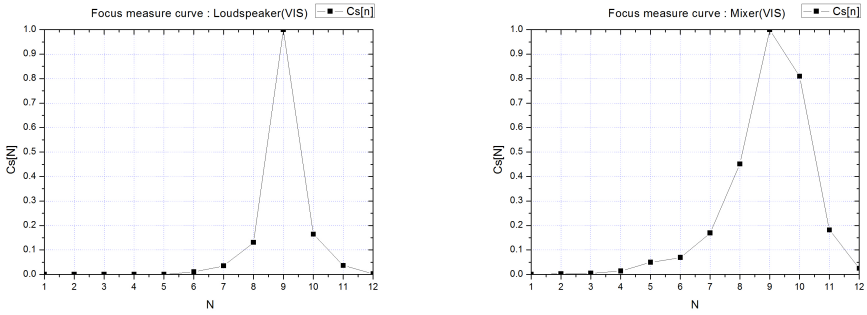


FIGURE 2.16: Specimen focus curves for 'Loudspeaker' and 'Mixer' image sets in visual spectrum.

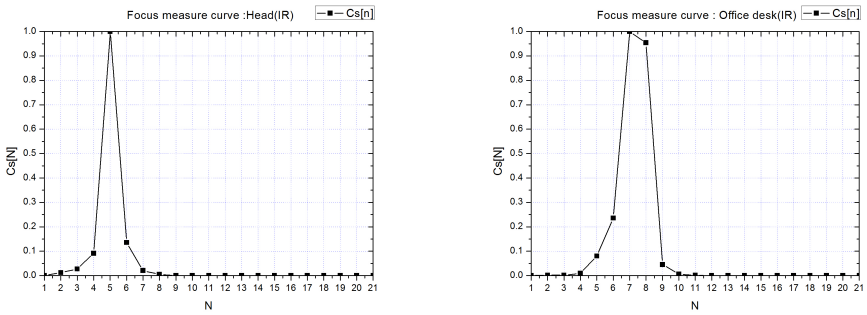


FIGURE 2.17: Specimen focus curves for 'Head' and 'Office desk' image sets in near-infrared spectrum.

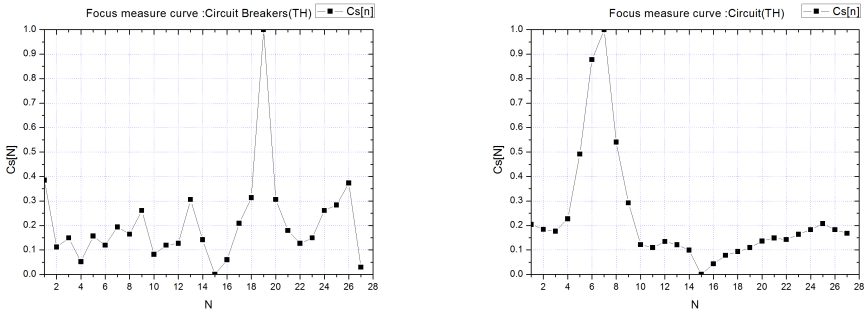


FIGURE 2.18: Specimen focus curves for 'Circuit breakers' and 'Circuit' image sets in thermal spectrum.

Average execution time of obtaining the focus measure using the proposed method is 2 sec. on a desktop PC with 3.4 GHz Intel Core CPU and 8 GB RAM. Our method tends to be slower compared to some of the other focus measures because we have to

TABLE 2.1: P and Q factor values of seven sets of images in VIS spectrum.

Spectrum	Object	Metric	Proposed method	FH	FAST	HL	EOL	SML	SF	P (Subjective)
VIS	Guitar	P	5	6	5	8	5	5	5	5
		Q	0.3333	0.2500	0.2500	0.2000	1.0000	0.5000	0.5000	-
	Head phones	P	7	5	8	0	7	7	7	7
		Q	0.3333	0.1667	0.3333	0.3333	0.5000	0.5000	0.5000	-
	Key board	P	5	4	3	0	5	5	5	5
		Q	0.2000	0.2000	0.1250	0.3333	0.3333	0.2000	0.1667	-
	Keys	P	3	0	2	7	2	2	2	2
		Q	0.3333	0.1667	0.2000	0.2500	0.5000	0.5000	0.5000	-
	Loud speaker	P	9	5	0	8	8	8	8	8
		Q	1.0000	0.3300	0.3333	0.1429	1.0000	1.0000	1.000	-
	Mixer	P	9	8	5	11	8	8	8	8
		Q	0.2500	0.5000	0.3300	0.1250	1.0000	1.0000	0.3300	-
	Sunglass	P	6	5	6	7	5	5	5	5
		Q	0.3333	0.5000	0.1250	0.3333	0.5000	0.2500	0.3333	-

TABLE 2.2: P and Q factor values of seven sets of images in NIR spectrum.

Spectrum	Object	Metric	Proposed method	FH	FAST	HL	EOL	SML	SF	P (Subjective)
NIR	Building	P	6	6	20	6	6	6	6	6
		Q	0.2000	0.0667	0.5000	0.2500	0.2500	0.2500	0.1667	-
	Car	P	6	6	4	20	4	4	4	7
		Q	0.2500	0.0667	0.2500	0.2500	0.5000	0.2500	0.2500	-
	Corridor	P	7	4	20	7	7	7	7	7
		Q	0.2000	0.1000	0.1000	0.2500	0.3333	0.1429	0.1429	-
	Head	P	5	4	19	2	4	4	4	4
		Q	1.0000	0.3333	0.1667	0.2000	1.0000	1.0000	1.0000	-
	Keyboard	P	5	4	16	0	4	4	4	4
		Q	1.0000	0.3300	0.3333	0.1429	1.0000	1.0000	1.0000	-
	Office Desk	P	7	6	7	10	6	6	6	6
		Q	0.5000	0.2000	0.3333	0.0909	1.0000	1.0000	1.0000	-
	Pens	P	8	7	18	3	7	7	7	7
		Q	1.0000	0.2500	0.2500	0.1429	1.0000	1.0000	1.0000	-

compute the local frequency map at thirteen different orientations. However, please note that Minhas et al. in [53] have reported an average execution time for their orientation-based focus measure to be 3.5 sec. for the same window size of 7×7 as ours.

TABLE 2.3: P and Q factor values of seven sets of images in TH spectrum.

Spectrum	Object	Metric	Proposed method	FH	FAST	HL	EOL	SML	SF	P (Subjective)
TH	Circuit Breakers	P	19	17	20	18	17	17	17	17
		Q	1.0000	0.5000	0.0769	0.2500	0.5000	0.0400	0.0476	-
	Building	P	25	25	3	23	12	12	25	25
		Q	1.0000	0.5000	0.0500	0.2000	0.1000	0.0476	1.0000	-
	Circuit	P	7	6	26	5	4	26	5	5
		Q	0.3333	0.3333	0.0900	0.1250	0.3333	0.0370	0.3333	-
	Engine	P	15	15	0	16	14	14	17	14
		Q	0.2500	0.5000	0.2000	0.2000	0.3333	0.0370	0.1429	-
	Printer	P	25	17	0	16	0	0	0	18
		Q	0.2000	0.3333	0.0909	0.2000	0.3333	0.0435	0.0500	-
	Server	P	20	21	3	18	20	6	20	20
		Q	0.3333	0.3333	0.1000	0.1667	1.0000	0.0435	1.0000	-
	Tube	P	6	19	0	20	0	2	0	20
		Q	0.1670	0.0500	0.0625	0.1429	0.1667	0.0714	0.0526	-

2.5.5 Performance analysis for SLF based fusion (First method)

In regards to multifocus fusion of images in the visual spectrum, we compare our method with the highly accurate multiresolution transform domain methods such as Discrete Wavelet Transform (DWT), Stationary Wavelet Transform (SWT), Curvelet Transform (CVT), Contourlet Transform (CT), Dual Tree Complex Wavelet Transform (DTCWT) and Non-Subsampled Contourlet Transform (NSCT) [37, 51, 77]. Since our fusion method is essentially based on interest point detection, we also compare our method with best performing Fast Hessian (FH) based fusion scheme from the same category [23]. For thermal fusion, we choose a recently reported EOL based method [52] for comparison. In addition, we also compare with FH based fusion scheme. Overall, we provide an extensive comparison with several recent and well-known spatial and transform domain based fusion methods.

Table 2.4 shows that the performances are improved with the inclusion of consistency verification (CV) for the proposed method. Fig. 2.19 qualitatively demonstrates the same results. Since only one image set pertaining to medical database is available, we just specify the result of the proposed method without any comparison. Next, in Table 2.5, we show that our method performs better compared to all the multiresolution transform based methods [37], in terms of a much higher MI, slightly higher Q_0 and comparable with $Q^{AB/f}$ (only marginally lower). Comparison to FH based fusion scheme reveals an improvement in terms of MI and $Q^{AB/f}$ values.

In VIS spectrum, for perceptual quality evaluation of fused images, we incorporate fused images obtained using FH, DWT and DTCWT methods (see Fig. 2.20-2.22). DWT is a basic method and performs moderately well while DTCWT is highly efficient [37]. The quality of fused images obtained using FH method are found to be inferior compared to our method. Some artifacts in form of halo effect are observed in case of DWT based fusion (shown using a red square in Fig 2.21). So, we can

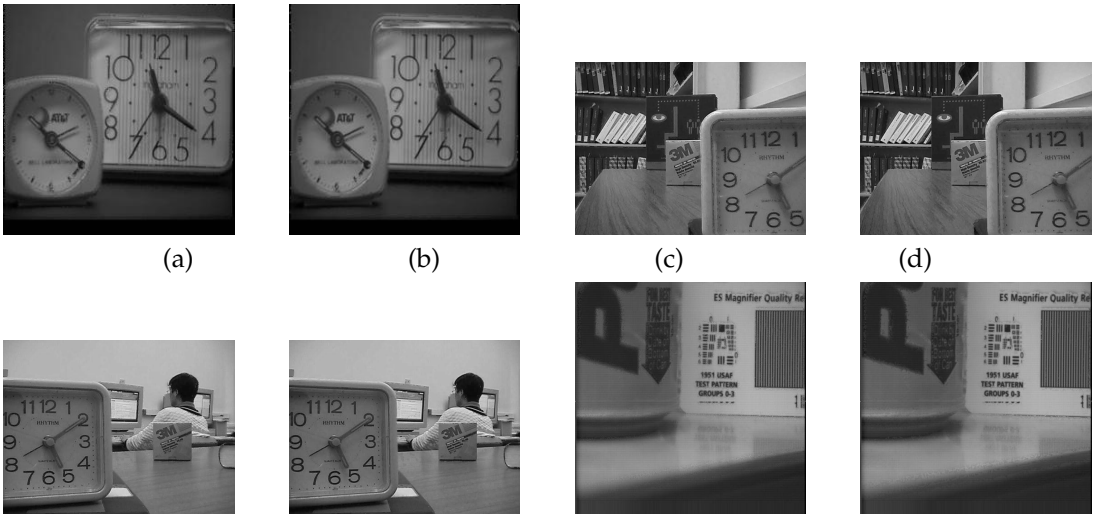
infer that the proposed method supersedes DWT based fusion. The perceptual quality of fused images obtained by DTCWT method is however comparable with the proposed method. But quantitatively in terms of metrics MI and Q_0 the proposed fusion method surpasses the DTCWT, while it is comparable in case of $Q^{AB/f}$. In the NIR spectrum, the visual quality of fused images obtained using the proposed method is comparable with that of the FH based method and is better than that of DWT and DTCWT based methods (see Fig. 2.23). For fusion in the thermal spectrum, the proposed scheme performs significantly better in comparison with the pixel-level weighted averaging method [52] and FH based scheme by yielding lower RMSE for all the five datasets. These results are specified in Table 2.6. In addition, we show the values obtained for CC and MAE in Table 2.7 which are quite promising. The fused results presented in Fig. 2.24 and 2.25 clearly indicate that the fusion scheme based on our focus measure gives high quality output.

TABLE 2.4: Fusion results of the proposed SLF based method: MI, $Q^{AB/f}$ and Q_0 values with and without Consistency Verification (CV).

Images	MI		$Q^{AB/f}$		Q_0	
	Without CV	**With CV**	**Without CV**	**With CV**	**Without CV**	**With CV**
Multifocus (VIS spectrum):						
Clock	8.5045	8.5563	0.6179	0.6701	0.9783	0.9785
Desk	7.9716	8.0246	0.5979	0.6697	0.9585	0.9586
Lab	8.3728	8.4551	0.6160	0.6838	0.9758	0.9759
Pepsi	8.2292	8.2621	0.6344	0.6838	0.9810	0.9810
Multifocus (NIR spectrum):						
Keyboard	7.9217	7.9342	0.6561	0.6853	0.9907	0.9908
Medical:						
CT-MR	7.0279	7.0278	0.6677	0.6870	0.5010	0.5028

TABLE 2.5: Multifocus image fusion by the proposed SLF based method: Performance comparison with (a) FH IPD based method, and (b) Best results of multi-resolution based fusion methods.

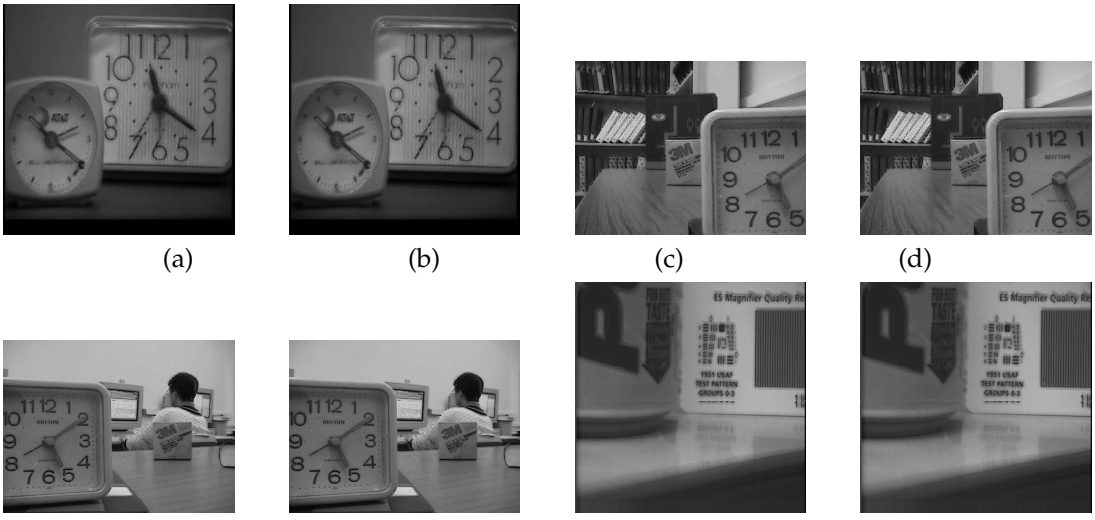
Spectrum	**Method**	MI	$Q^{AB/f}$	Q_0
VIS spectrum	Proposed SLF method	**8.3245**	**0.6768**	**0.9735**
	FH	8.2493	0.5804	0.9746
	DWT	2.4126	0.6866	0.7206
	SWT	2.4510	0.7140	0.7555
	DTCWT	2.4814	0.7231	0.7650
	CVT	2.4387	0.7075	0.7421
	CT	2.3978	0.6700	0.7076
	NSCT	2.4804	0.7219	0.7799
NIR Spectrum	Proposed SLF method	**7.9342**	**0.6853**	**0.9908**
	FH	8.0198	0.7105	0.9912
	DWT	5.9485	0.5135	0.9061
	DTCWT	7.3575	0.7082	0.9902



(a) (b) (c) (d)

(e) (f) (g) (h)

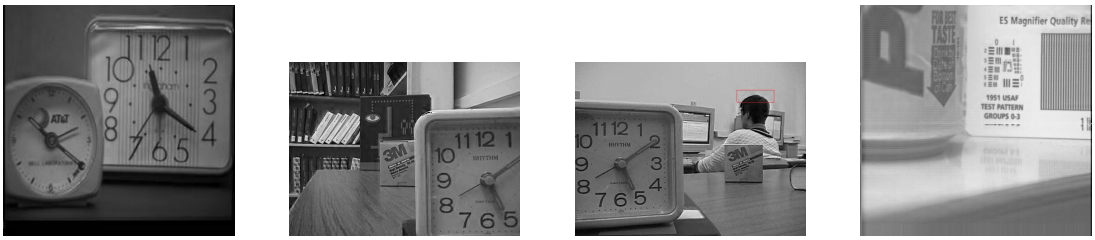
FIGURE 2.19: Fused images obtained by the proposed SLF based method in the VIS spectrum for four datasets performed without consistency verification (FI) and with consistency verification (FI-CV). (a) 'Clock' FI , (b) 'Clock' FI-CV ; (c) 'Desk' FI, (d) 'Desk' FI-CV; (e) 'Lab' FI, (f) 'Lab' FI-CV; (g) 'Pepsi' FI, (h) 'Pepsi' FI-CV.



(a) (b) (c) (d)

(e) (f) (g) (h)

FIGURE 2.20: Fused images obtained using the Fast Hessian (FH) IPD based method in the VIS spectrum for four datasets performed without consistency verification (FI) and with consistency verification (FI-CV). (a) 'Clock' FI , (b) 'Clock' FI-CV ; (c) 'Desk' FI, (d) 'Desk' FI-CV; (e) 'Lab' FI, (f) 'Lab' FI-CV; (g) 'Pepsi' FI, (h) 'Pepsi' FI-CV.

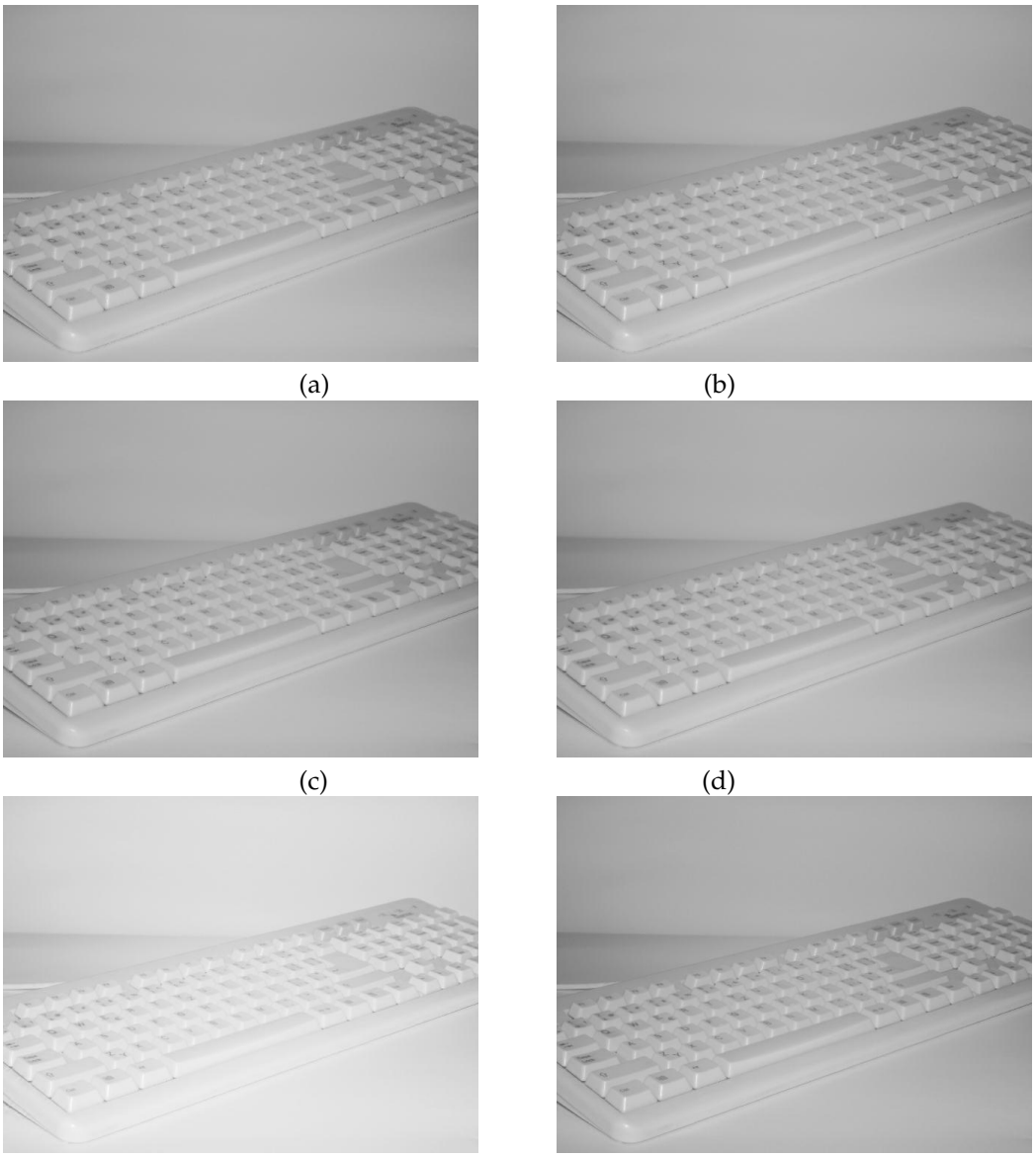


(a) (b) (c) (d)

FIGURE 2.21: Fused images obtained by DWT method in the VIS spectrum for four datasets. (a) 'Clock', (b) 'Desk', (c) 'Lab', (d) 'Pepsi'.

(a) (b) (c) (d)

FIGURE 2.22: Fused images obtained by DTCWT method in the VIS spectrum for four datasets. (a) 'Clock', (b) 'Desk', (c) 'Lab', (d) 'Pepsi'.



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

FIGURE 2.23: Fused images obtained by the proposed SLF based method in NIR spectrum performed without consistency verification (FI) and with consistency verification (FI-CV). (a) proposed method FI (b) proposed method FI-CV , (c) FH IPD based method FI , (d) FH IPD based method FI-CV (e) DWT method (f) DTCWT method.

TABLE 2.6: TH multifocus image fusion with reduced dataset by the proposed SLF based method: RMSE.

Image set	**Pixel level Weighted averaging method on EOL AL Fusion [52]**	**FH method**	**Proposed SLF method**
Mobile-RS232	0.1803	0.0183	0.0172
Bulbs set 1	0.1999	0.0307	0.0184
Bulbs set 2	0.1342	0.0293	0.0160
Bulbs set 3	0.2648	0.0541	0.0313
Bulbs set 4	0.3307	0.0589	0.0368

TABLE 2.7: TH multifocus image fusion with reduced dataset by the proposed SLF based method: CC and MAE.

Image set	**FH method**		**Proposed SLF method**	
	CC	MAE	CC	MAE
Mobile-RS232	0.9933	0.0077	0.9944	0.0078
Bulbs set 1	0.9942	0.0097	0.9979	0.0075
Bulbs set 2	0.9891	0.0167	0.9969	0.0097
Bulbs set 3	0.9832	0.0250	0.9946	0.0210
Bulbs set 4	0.9821	0.0348	0.9932	0.0309

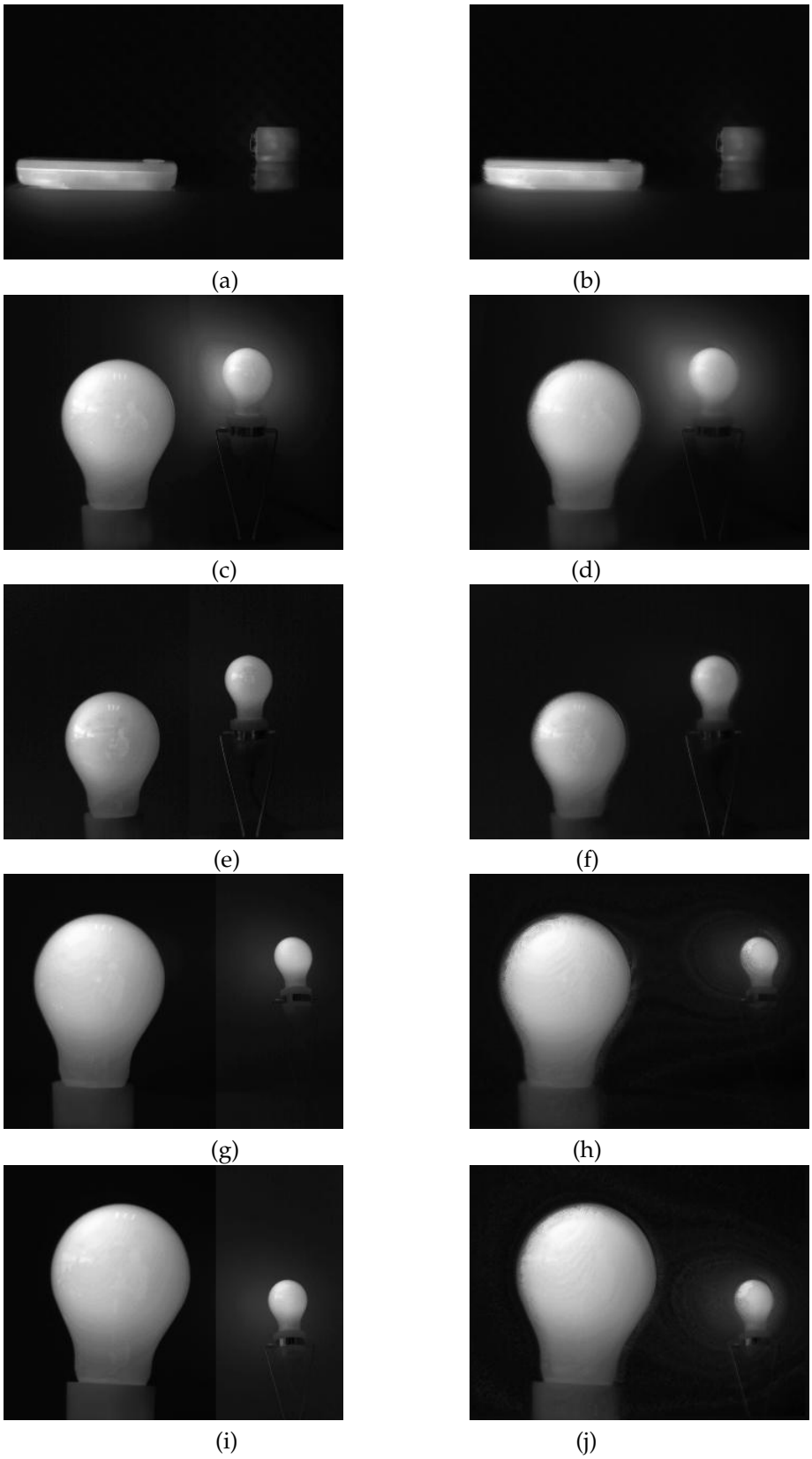
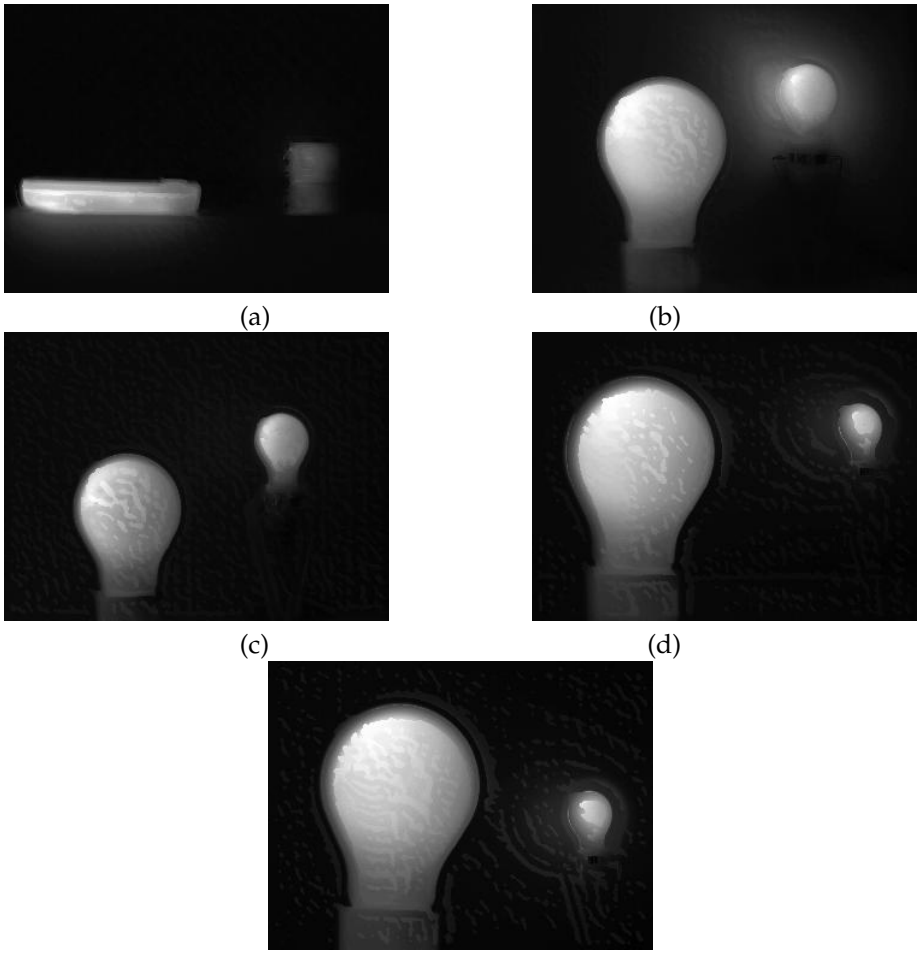
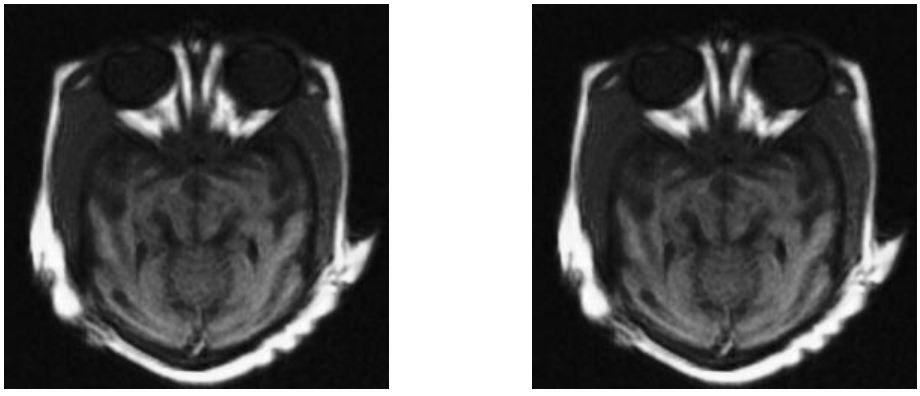


FIGURE 2.24: Ground truth (GT) and fused images (FI) obtained using the proposed SLF method in the TH spectrum for five datasets. (a) Mobile_RS232 GT, (b) Mobile_RS232 FI; (c) Bulbs Set 1 GT, (d) Bulbs Set 1 FI; (e) Bulbs Set 2 GT, (f) Bulbs Set 2 FI; (g) Bulbs Set 3 GT, (h) Bulbs Set 3 FI; (i) Bulbs Set 4 GT, (j) Bulbs Set 4 FI.



(a) (b) (c) (d) (e)

FIGURE 2.25: Fused images obtained using the Fast Hessian (FH) IPD based method in the TH spectrum for five datasets. (a) Mobile_RS232 , (b) Bulbs Set 1 ; (c) Bulbs Set 2, (d) Bulbs Set 3; (e) Bulbs Set 4.



(a) (b)

FIGURE 2.26: Fusion of CT and MRI modality images obtained by the proposed SLF based method: (a) Without consistency verification (FI), (b) With consistency verification (FI-CV).

2.5.6 Performance analysis for GSLF and improved saliency model based fusion (GSLF-IS/Second method)

In the VIS and NIR spectra, we compare our results with some spatial as well as some transform domain based approaches [37] as in previous section 2.5.5. Fast Hessian (FH) being one of the best in the interest point detection (IPD) based category, is also chosen for comparison. We also compare the results with our previous steerable local frequency (SLF) based method results, see section 2.5.5. Results of the proposed method in the VIS spectrum at an intermediate stage, i.e., after only guided filtering of SLF and without saliency (GSLF) are also incorporated. We use the recently reported EOL based method [52] in addition to FH and SLF method for comparisons in the TH spectrum. Table 2.8 shows quantitative results of the proposed method in the VIS and NIR spectra. Corresponding fused results are shown in Fig. 6 and 7. Perceptual quality of the fused images is superior as it shows no inconsistencies or artifacts in the form of halo effects, illumination changes and contrast reduction. Table 2.9 shows the quantitative comparison of our method with other methods. In the VIS spectrum, the proposed method performs better than all the multiresolution based methods and the SLF as well as the GSLF with significant improvement in MI, $Q^{AB/f}$ and Q_0. In comparison to the FH IPD based method, the proposed method supersedes in terms of MI, $Q^{AB/f}$ and marginally looses in terms of Q_0. In the NIR spectrum, the proposed method shows consistently good performance over the SLF, FH, DWT and DTCWT methods. The quantitative results in TH spectrum are shown in Table 2.10 and 2.11. The proposed fusion method outperforms the pixel level weighted averaging with EOL activity level based fusion method [52] with high margin in terms of RMSE. Our method also shows an improvement over the SLF and FH IPD based methods in terms of $RMSE$, CC and MAE. The proposed method yielded better results as compared to its competitors due to the use of guided filter, steerable local frequency maps and the improved saliency maps. Fig. 2.29 (a-e) shows the fused resultant images in the TH spectrum for five image sets (Set1-Set5).

TABLE 2.8: Fusion results of the GSLF and improved saliency based method (GSLF-IS) in VIS and NIR spectra: MI, $Q^{AB/f}$ and Q_0 values with Consistency Verification (CV).

Images	MI	$Q^{AB/f}$	Q_0
VIS spectrum			
Clock	8.7171	0.7405	0.9786
Desk	8.3108	0.7313	0.9588
Lab	8.5689	0.7437	0.9761
Pepsi	8.8772	0.7849	0.9812
NIR spectrum			
Keyboard	8.1360	0.7580	0.9980

TABLE 2.9: Multifocus image fusion results of GSLF and improved saliency based method (GSLF-IS): Performance comparison in VIS and NIR spectra

Spectrum	**Method**	MI	$Q^{AB/f}$	Q_0
VIS spectrum	**GSLF-IS method**	**8.6185**	**0.7501**	**0.9737**
	GSLF	8.5930	0.7496	0.9737
	SLF method	8.3245	0.6768	0.9735
	FH	8.2493	0.5804	0.9746
	DWT	2.4126	0.6866	0.7206
	SWT	2.4510	0.7140	0.7555
	DTCWT	2.4814	0.7231	0.7650
	CVT	2.4387	0.7075	0.7421
	CT	2.3978	0.6700	0.7076
	NSCT	2.4804	0.7219	0.7799
Near-infrared (NIR) spectrum	**GSLF-IS method**	**8.1360**	**0.7580**	**0.9980**
	SLF method	7.9342	0.6853	0.9908
	FH	8.0198	0.7105	0.9912
	DWT	5.9485	0.5135	0.9061
	DTCWT	7.3575	0.7082	0.9902

Visual inspection of the fused images in comparison with the ground truth clearly demonstrates the superior performance of the proposed fusion method.

(a) (b) (c) (d)

FIGURE 2.27: Fused images obtained using the GSLF and improved saliency based method (GSLF-IS) in the VIS spectrum for four datasets with consistency verification (CV): (a) Clock, (b) Desk, (c) Lab, (d) Pepsi.

TABLE 2.10: TH multifocus image fusion with reduced dataset by the proposed GSLF-IS method: *RMSE*.

Image Set	**Pixel level Weighted averaging based on EOL AL Fusion** [15]	**FH**	**SLF method**	**GSLF-IS method**
Mobile-RS232	0.1803	0.0183	0.0172	**0.0159**
Bulbs set 1	0.1999	0.0307	0.0184	**0.0150**
Bulbs set 2	0.1342	0.0293	0.0160	**0.0118**
Bulbs set 3	0.2648	0.0541	0.0313	**0.0210**
Bulbs set	0.3307	0.0589	0.0368	**0.0265**

TABLE 2.11: TH multifocus image fusion with reduced dataset by the proposed GSLF-IS method: *CC* and *MAE*.

Image set	**FH**		**SLF method**		**GSLF-IS method**	
	CC	*MAE*	*CC*	*MAE*	*CC*	*MAE*
Mobile-RS232	0.9933	0.0077	0.9944	0.0078	**0.9951**	**0.0075**
Bulbs set 1	0.9942	0.0097	0.9979	0.0075	**0.9986**	**0.0063**
Bulbs set 2	0.9891	0.0167	0.9969	0.0097	**0.9983**	**0.0072**
Bulbs set 3	0.9832	0.0250	0.9946	0.0210	**0.9976**	**0.0146**
Bulbs set 4	0.9821	0.0348	0.9932	0.0309	**0.9965**	**0.0219**

FIGURE 2.28: Fused images obtained by the GSLF-IS based method in the NIR spectrum: Keyboard set.

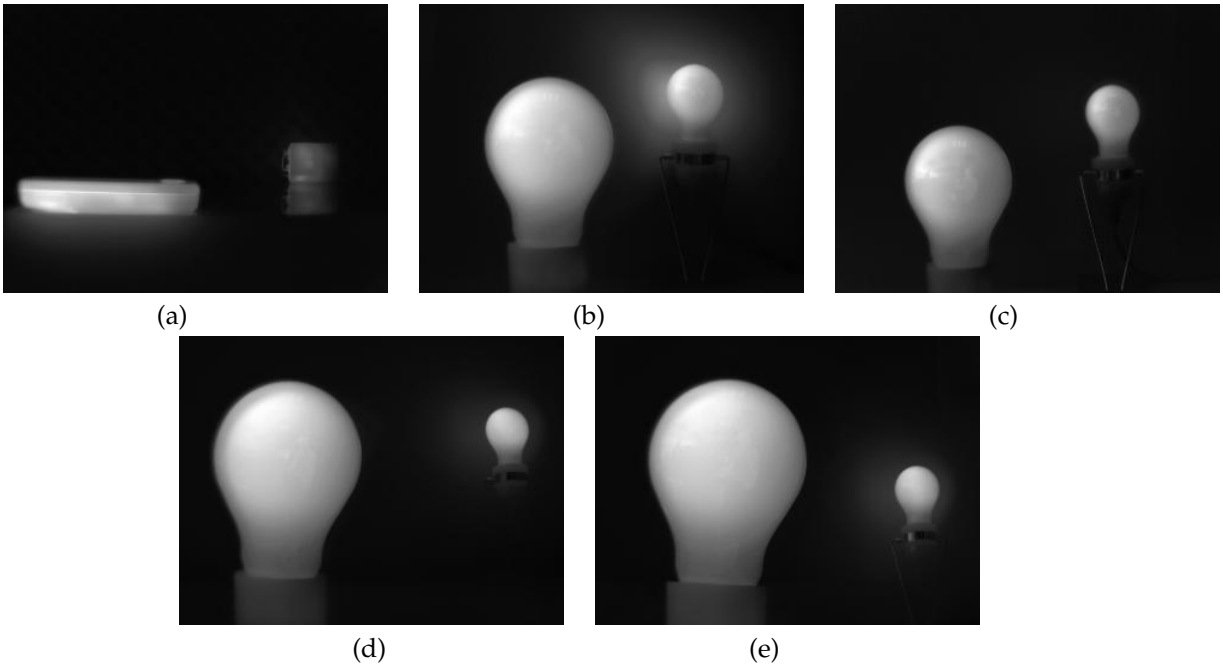


FIGURE 2.29: Fused images obtained using the GSLF-IS based method in the TH spectrum: (a) Mobile phone and RS232 Set; (b) Bulbs Set 1; (c) Bulbs Set 2; (d) Bulbs Set 3; (e) Bulbs Set 4.

2.6 Discussions

In our first work, we present a new focus measure based on steerable local frequency based interest point detection. The proposed focus measure is shown to perform well in different spectra. Better performance of the proposed focus measure is due to the use of orientation selective local frequency in the source images. We further demonstrated that the proposed focus measure improves mutispectral multifocus fusion. In the visual spectrum, our fusion scheme outperforms some of the robust and efficient multiresolution transform based methods in addition to some IPD based approaches. In the near-infrared spectrum the proposed fusion method offers a decent performance in comparison with the spatial and transform domain based approaches. In the thermal spectrum, the results show significant improvement over previously reported results.

To achieve very consistence and better fusion performance, in the next work, we make innovative use of guided filtering and an improved saliency model. Superior fusion results are achieved by combining guided steerable local frequency maps with the saliency maps over all spectra.

Chapter 3

Multispectral Causal Video Fusion

In this chapter, in first part, we discuss a novel proposed framework for semantic segmentation of causal video using superseeds and graph matching. Which performs well compared to recently reported works. In the second part, we discuss the proposed superpixel based causal multispectral video fusion algorithm suitable for real-time surveillance tasks. Here we make use of the techniques we developed in causal video segmentation in the first part. We develop new superpixel level spatial and temporal saliency models. Novel superpixel level multiple fusion rules are also designed to obtain the fused output. Comprehensive comparisons with several existing works clearly indicate the benefit of our solution.

3.1 Introduction

Video surveillance systems have become extremely important due to increasing security concerns. Single modality (spectrum) surveillance systems [78] work well in controlled conditions for civilian, military, and, remote sensing applications. However, these systems often fail in cases of low illumination, shadowing, smoke, dust, unstable background, and camouflage. Recent developments in sensor technologies have led to the popularity of multispectral surveillance systems which can perform better in such adverse situations. The multispectral surveillance systems make use of fusion of videos from different spectra like visible and infrared. The fused video can describe a scene more accurately and precisely as compared to any of the individual modalities [79] applied in isolation. Any multispectral video fusion system should

have the following characteristics [3] - (i) it should preserve all relevant information of the input video pairs, (ii) it should not introduce any artifacts and inconsistencies (false information); (iii) it should be temporally stable and consistent; and (iv) it should be shift-invariant. In addition, the algorithm should have a very low execution time in order to be used for any real-time application. The fused video can be further analyzed for various important surveillance tasks like anomalous event detection [8, 80] and person re-identification [9], fire detection [81], video copy detection [82] and video search [83].

Early multispectral video fusion methods essentially followed an independent frame by frame fusion approach [3, 4, 10, 11]. Such strategies suffer from temporal instability and inconsistencies. To overcome this problem, some of the later methods utilized the information from the adjacent past and future frames to fuse the current frame [12–14]. However, for real-time applications, the future frames may not be available in the system at the time of processing the current frame. Such a system is termed as a causal system. So, in causal video fusion, one can only make use of the past frames to fuse the current frame. Majority of the video fusion approaches reported in the literature are non-causal in nature. Hence, the main motivation behind this work is to propose a causal video fusion algorithm which can be applied for real-time video surveillance applications. Majority of non-causal video fusion methods are transform domain based [12–14, 24, 25]. The transform domain approaches suffer from inherent information loss due to approximation issues in implementation. Furthermore, these methods could sometimes be computationally quite expensive due to the use of higher scales, necessary to achieve robust performance. Spatial domain processing is often more accurate [84]. However, conventional pixel level spatial domain processing of high volume data could seriously restrict its use for real-time applications.

It has been observed that the high fusion performance can be achieved if we first identify spatial and temporal information and then merge these using appropriate

fusion strategies. So, as a first step, to segment video frames into meaningful regions, we propose a novel causal video segmentation using superseeds and graph matching. We first employ Simple Linear Iterative Clustering (SLIC) for the extraction of superpixels from video frames in a causal manner. A set of superseeds is chosen from the superpixels in each frame using color and texture based spatial affinity measure. Temporal coherence is ensured through propagation of labels of the superseeds across each pair of adjacent frames. A graph matching procedure based on comparison of the eigenvalues of graph Laplacians is employed for label propagation. Watershed algorithm is applied finally to label the remaining pixels to achieve final segmentation.

In the second step, we propose a novel superpixel based framework for causal multispectral (visible and infrared/thermal infrared) video fusion (CMVF) which can be very useful for real-time video surveillance applications. Major step of the proposed framework is to segment input video frames into four types of regions- uniform, spatially salient, temporally salient and spatio-temporally salient regions which is based on our causal video segmentation algorithm. The proposed approach consists of three stages- pre-processing, saliency detection and fusion. In the pre-processing stage, we extract superpixels from the visible (VIS) and infrared or thermal infrared (IR) video frames. In the second stage, we obtain spatial and temporal saliency maps for these frames at the superpixel level. To build spatial saliency maps, we use superpixel level color and texture information. Temporal saliency detection is based on superpixel level direct frame difference (SDFD) and local region graph matching between current and previous frame. In the third and final stage, we propose superpixel level fusion rules.

3.2 Related works

We first mention some of the early works for multispectral video fusion based on extension of multisensor image fusion. In [3], Rockinger suggested fusion of VIS and TH-IR spectrum videos for object detection and tracking for surveillance applications. Bennet et al. [10] developed a scheme for enhancing underexposed visible spectrum video by fusing it with co-registered simultaneously captured video from Short wave IR or Near IR for video enhancement. Denman et al. [4] demonstrated how the fusion of simultaneously captured multiple modality (multi-spectra) videos can enhance the performance of surveillance systems. Rasmussen et al. [11] demonstrated, how fusion of VIS and IR videos helps wilderness search and rescue groups. In [24], Dixon presented an analogy between the multispectral video fusion in visible and infrared domain with the visual system of rattlesnakes. In [85], Torabi et al. proposed an integrated framework for TH-IR and visible image registration, fusion and tracking for video surveillance applications. Recently, Pillai and Swamy [86] proposed a frame by frame real-time video fusion algorithm for camouflaged target detection. Note that the above works suffer from temporal instability and inconsistencies as they followed a frame by frame approach and exploited only spatial information. In addition, we observed that execution time of [86] is still somewhat high for real-time applications.

Now, we discuss some video fusion approaches which make use of temporal information. Zhang et al. in [12] proposed a framework for multispectral video fusion based on the motion selective multiscale analysis tool using 3D surfacelet transform. However, the method does not work well on the videos with highly dynamic background. In a second work [13], they proposed an algorithm based on spatial-temporal saliency detection with 3D uniform Curvelet transform (3D-UDCT). This method, though more robust than the previous method, suffers from poor execution time. So, in [14], they proposed another video fusion algorithm based on the 3D surfacelet (3D-ST) transform and higher order singular value decomposition (HOSVD).

This method suppresses unwanted scene noise and has a very low execution time. In another recent work, Xu et al. [25] proposed a method for fusing videos from the visible and the IR spectra based on motion compensated wavelet transform. Note that all these transform domain methods essentially work in a non-causal manner, i.e., while processing the current frame, they made use of both the past and future frames. In addition, the transform domain approaches suffer from inherent information loss due to approximation issues in implementation.

To the best of our knowledge, there is no causal multispectral video fusion algorithm available till date. We propose a spatial domain superpixel level causal multispectral multispectral video fusion algorithm. The main contributions of this work are highlighted below:

1. Superpixel level accurate spatio-temporal saliency detection. The spatial saliency detection is based on a combination of color and texture measures. Temporal saliency detection is achieved from superpixel level direct frame difference (SDFD) and local region graph matching.

2. Superpixels are grouped into four different categories based on their saliency values. Appropriate fusion rules are designed at the superpixel level yielding a highly accurate fused video.

3. Extremely fast execution time making the method a highly suitable candidate for real-time surveillance applications. This has been possible due to superpixel level processing.

3.3 Superpixel Extraction

Superpixel extraction significantly reduces computational complexity in video segmentation algorithms [87, 88]. We use the SLIC algorithm [89] for the extraction of

superpixels in each frame of a causal video. So, we can write:

$$I_{t,SLIC} = SLIC(I_t, k) \tag{3.1}$$

where I_t is the current frame and $I_{t,SLIC}$ is the frame with extracted superpixels. The inputs to SLIC are the current frame I_t and the desired number of superpixels k. The CIELAB color space is used for clustering color images. In an initialization step, k initial cluster centers $C_i, i = 1, ..., k$ are sampled on a regular grid with spacing S pixels. Hence, we can write:

$$C_i = [l_i, a_i, b_i, x_i, y_i]^T \tag{3.2}$$

$$S = \sqrt{\frac{N}{k}} \tag{3.3}$$

where N is the number of pixels in the image. The seed centers C_i are moved to locations with lowest gradient position in 3×3 neighborhood. Then, each pixel i is associated with the nearest cluster center. Limiting the size of search region to $2S \times 2S$ around the center significantly reduces the computation compared to the k-means clustering. A new distance measure D which is a combination of color distance (d_c) in CIELAB space and spatial distance (d_s) is used for that purpose. The update step then adjusts each cluster center to be the mean $[l, a, b, x, y]^T$ vector of all the pixels of that cluster. For our work. we find 10 iterations to be sufficient to reach the convergence.

3.4 Proposed Causal Video Segmentation

Video segmentation [90–92] aims at grouping pixels into meaningful spatio-temporal regions that exhibits coherence in appearance and motion. The problem of video segmentation [93–95] becomes extremely challenging due to size of the input, camera motion, occlusions, non-rigid object motion, and uneven illumination. Video segmentation techniques can be classified into non-causal (off-line) and causal (on-line) categories. While non-causal segmentation techniques make use of both the past and future video frames, causal segmentation approaches rely only on the past frames. For some recently reported causal video segmentation works, please see [87, 88, 96, 97]. Some of these algorithms employ superpixels to reduce computational complexity and to achieve powerful within-frame representation [87, 88]. The method in [97] does not guarantee temporal consistency. Miksik et al. [96] performs semantic segmentation using optical flow to ensure temporal consistency. But, complexity of pixel-level optical flow computation poses a serious constraint for its use in real-time applications. Couprie et al. [87] proposed an efficient causal graph-based video segmentation method using minimum spanning tree. However, the method uses some heuristics in both the pre and post processing stages. We propose a novel framework for semantic segmentation of causal video using superseeds and local graph matching [98].

The proposed framework is illustrated in Fig. 3.1 as shown below.

SLIC [89] is applied for the generation of superpixels in each frame of a causal video. As a part of the initialization step, we apply the DBSCAN [99] (Density Based Spatial Clustering of Applications with Noise) method with some modifications resulting from our spatial consistency measure to achieve the final segmentation of the first frame. Some representative superpixels are then chosen using the above spatial affinity measure. We deem the centers of such superpixels as superseeds. Labels of these superseeds are propagated to the current frame from the previous frame by

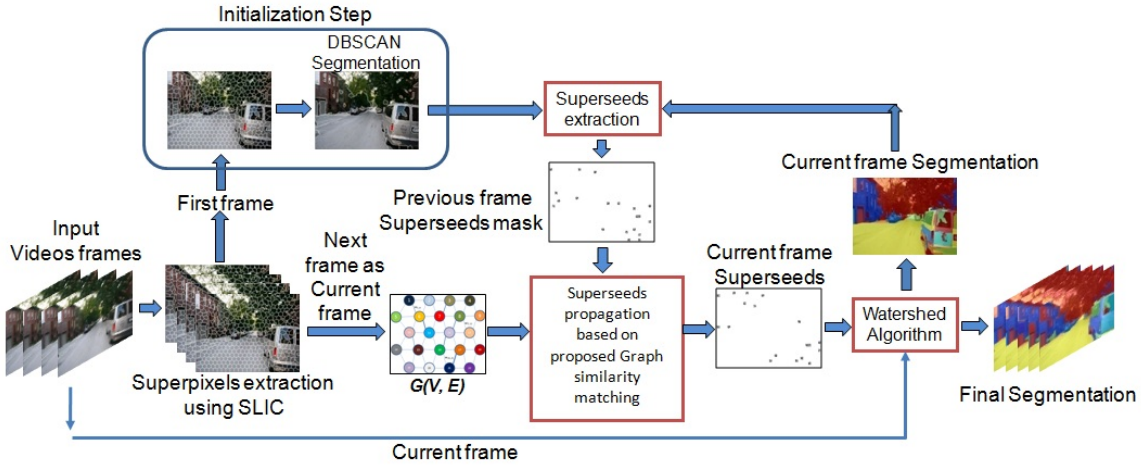


FIGURE 3.1: Schematic: Proposed causal video segmentation.

using local graph matching. Entries and exits are also handled efficiently to achieve temporal consistency. Watershed is applied to label the remaining pixels (other than the superseeds) to achieve complete segmentation of the current frame.

3.4.1 Spatial saliency measure

A hexagonal neighborhood graph $G = (V, E)$ is constructed with the extracted superpixels as the nodes using hexagonal grid as suggested by http://www.csse.uwa.edu.au/pk/research/matlabfns/Spatial/slic.m. This is shown in Fig. 3.2. The spatial affinity between two superpixels S_i and S_j is captured by the edge weights ω_{ij}. Color and texture information are used to compute these edge weights. For the color information, intersection (minimum) between cumulative color histograms of two superpixels under consideration is employed as a measure. This is given by:

$$c_{ij} = N\left[Hist(S_i) \cap Hist(S_j)\right] \tag{3.4}$$

Here, $Hist(\cdot)$ represents the cumulative color histogram of a superpixel. N is the normalization constant, set equal to $1/max(c_{ij})$. The larger the value of c_{ij}, the higher is the color affinity between the superpixels S_i and S_j. For the texture information measure, we use a gray-scale local binary pattern (LBP) [100] based measure. The $LBP_{P,R}$

number characterizes the local image structure and can be computed as follows:

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p \tag{3.5}$$

where p are the pixel within a circular neighborhood of radius R of the center pixel c. And g_p and g_c represents corresponding pixel intensities. We have taken P=8 and R=1 for our problem. The function s is given by:

$$s(x) = \begin{cases} 1 & \text{if} \quad x \geq 0 \\ 0 & otherwise \end{cases} \tag{3.6}$$

The above LBP number is computed for every pixel in a superpixel. We can write $\overline{ST_i}$, as a texture measure of the superpixel S_i, given by a joint vector:

$$\overline{ST_i} = \bigcup_{n=1}^{|S_i|} LBP_{P,R,n} \tag{3.7}$$

where n is a pixel in S_i. Similarly we can have $\overline{ST_j}$ for the superpixel S_j. The normalized texture affinity measure t_{ij} between two superpixels S_i and S_j is given by:

$$t_{ij} = 1 - \frac{W_H(\overline{ST_i} \bigoplus \overline{ST_j})}{max_{\forall i,j}[W_H(\overline{ST_i} \bigoplus \overline{ST_j})]} \tag{3.8}$$

Where $\overline{ST_j}$ is truncated to the length of $\overline{ST_i}$ and W_H is the Hamming weight function on binary vectors. Larger value of t_{ij} indicates higher texture affinity. Finally, we present the proposed spatial affinity measure between the superpixels S_i and S_j as:

$$\omega_{ij} = c_{ij} \times t_{ij} \tag{3.9}$$

Note that $\omega_{ij} \in [0\,1]$.

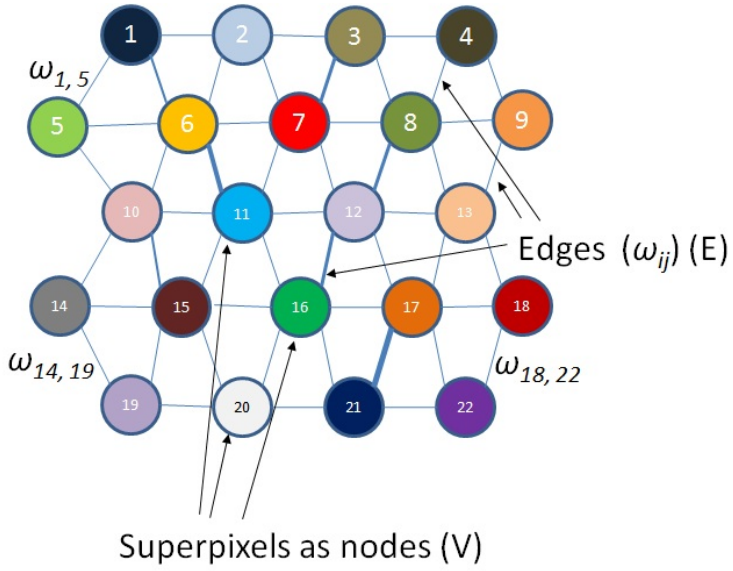


FIGURE 3.2: Superpixel neighborhood graph

3.4.2 Label propagation using graph similarity

We now mention the various steps linked with propagation of labels from the previous frame to the current frame. These steps are discussed below:

3.4.2.1 Selection of superseeds

In the initialization step, only the first frame is segmented by the modified DBSCAN [99] using the above spatial affinity measure. Each segment consists of multiple superpixels and we discard those segments which have less than two superpixels. The geometric centers of the remaining segments are extracted and treated as superseeds.

3.4.2.2 Local graph matching

Local region graphs are constructed surrounding each superseed in the previous frame and surrounding corresponding pixels (having same spatial locations as that of the superseeds in the previous frame)in the current frame. This is illustrated in Fig. 3.3. These two graphs are compared to propagate the label from the previous frame to the current frame. Let $G1(V1, E1)$ corresponds to the local region graph sur-

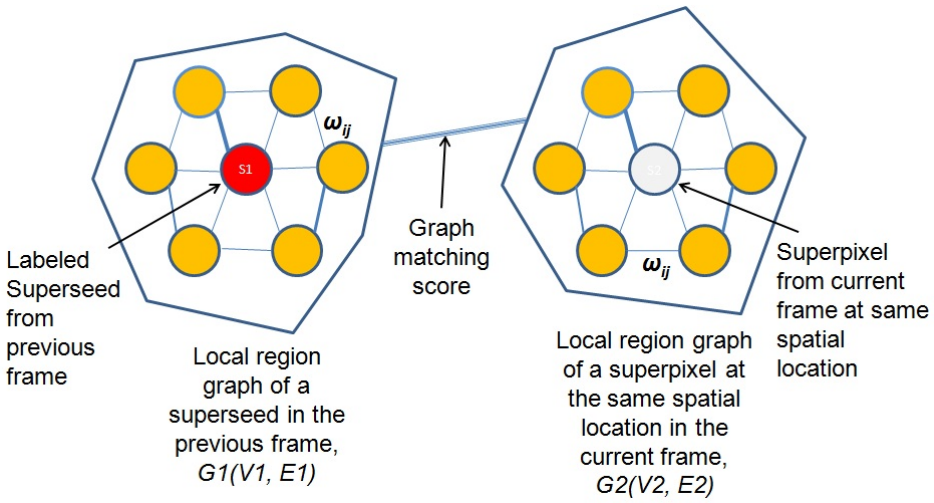


FIGURE 3.3: Local graph similarity matching

rounding a superseed in the previous frame. Similarly, let $G2(V2, E2)$ corresponds to the local region graph surrounding the pixel with same spatial location (as that of the superseed in the previous frame) in the current frame. We use graph Laplacian's eigenvalue-based score for matching [101]. Let $A1$ and $A2$ be the adjacency matrices, $D1$ and $D2$ be the diagonal matrices and $L1$ and $L2$ be the Laplacian matrices of the graphs $G1$ and $G2$ respectively. Then, we can write:

$$L1 = D1 - A1 \tag{3.10}$$

$$L2 = D2 - A2 \tag{3.11}$$

We use the similarity matching score $Sim_{G1,G2}$ between $G1$ and $G2$ by computing the top k eigenvalues of Laplacians $L1$ and $L2$, that contain 90% of energy, as given by:

$$Sim_{G1,G2} = \sum_{i=1}^{k} (\lambda_{1i} - \lambda_{2i})^2 \tag{3.12}$$

where k is chosen as shown below:

$$\min_j \left(\frac{\sum_{i=1}^{k} \lambda_{ji}}{\sum_{i=1}^{n} \lambda_{ji}} > 0.9 \right) \tag{3.13}$$

Low values of $Sim_{G1,G2}$ indicate that the graphs are very similar and vice-versa.

3.4.2.3 Temporal Consistency and Label Propagation

If the matching score (see equation 3.12) is less than an experimentally chosen threshold (T_1), then the two co-located regions under consideration have temporal coherence. So, we simply copy the label of the superseed of the previous frame to the next frame. If this score is higher, then there is no such temporal consistency between the two corresponding regions. This may occur due to an exit or a new entry in the current frame. To further differentiate between these two situations, we check the spatial affinity (ω_{ij}) of the superpixel in the current frame with its neighbors in the local region graph. If the spatial affinity is more than an experimentally chosen threshold (T_2), it signifies an exit and no new label is required in that case. If the spatial affinity is less, it signifies an entry and we assign a new label to the superpixel in the current frame. In this manner, we ensure temporal coherence between each successive pair of frames under different situations (with or without entry and/or exit).

3.4.3 Watershed for final segmentation

We next employ the sequential unordered watershed algorithm with respect to topographical distance function [102], derived from the shortest path algorithm, to label the remaining pixels in the current frame to achieve the final segmentation. The basics of watershed transform following [102, 103] is included for the sake of completeness. Let f be a gray value of the morphologically processed input frame(image). The lower slope $LS(p)$ at pixel p is defined as the maximal slope linking p to any of its neighbors of lower altitude. Thus,

$$LS(p) = \max_{q \in N_G(p) \cup \{q\}} \left(\frac{f(p) - f(q)}{d(p,q)} \right) \tag{3.14}$$

where $N_{G(p)}$ is the set of neighbors of pixel p on the grid graph $G = (V, E)$ built on f and $d(p, q)$ is the distance associated with the edge (p, q). The cost of walking from a pixel p to its neighboring pixel q is defined as:

$$cost(p,q) = \begin{cases} LS(p) \cdot d(p,q) & if \quad f(p) > f(q) \\ LS(q) \cdot d(p,q) & if \quad f(p) < f(q) \\ \frac{1}{2}\left(LS(p) + LS(q)\right) \cdot d(p,q) & if \quad f(p) = f(q) \end{cases} \tag{3.15}$$

The topographical distance along a path π between p and q is defined as:

$$T_f^{\pi}(p,q) = \sum_{i=0}^{l-1} d(p_i, p_{i+1}) \cdot cost(p_i, p_{i+1}) \tag{3.16}$$

The topographical distance between p and q is the minimum of the topographical distances along all paths between p and q and is defined as:

$$T_f(p,q) = \min_{\pi \in [p \longrightarrow q]} T_f^{\pi}(p,q) \tag{3.17}$$

Let $(m_i)_{i \in I}$ be the collection of minima (markers) of f. The catchment basins $CB(m_i)$ of f correspond to a minimum m_i is defined as the basin of the lower completion of f:

$$CB(m_i) = \{p \in D \mid \forall j \in I \backslash \{i\} : f^*(m_i) + T_{f^*}(p, m_i) < f^*(m_j) + T_{f^*}(p, m_j)\} \tag{3.18}$$

where f^* is the lower completion of f. The watershed of f with 2D grid D are the points which do not belong to any catchment basin and is defined in the following manner:

$$Wshed(f) = D \cap (\cup_{i \in I} \cdot CB(m_i))^c \tag{3.19}$$

The superseeds generated in the earlier stage of our solution pipeline act as the markers (regional minima). Thus construction of the catchment basins (segments) of the frame becomes a problem of finding a path of minimal cost between each pixel and a

marker (regional minima). Note that for the second frame onwards, the watershed-based final segmentation provides the labels of the superpixels in the current frame. We then propagate the labels of the superseeds in the current frame to the next frame using the graph matching technique.

3.4.4 Experimental results

We have implemented the proposed method in MATLAB $R2013b$ environment on a desktop PC with $3.4GHz$ Intel Core i7 CPU with $8GB$ RAM. SLIC for superpixels extraction is used from [89] and DBSCAN from [99]. The average execution time of the proposed method is 3.5 sec. out of which SLIC itself takes 3 sec. The values of the thresholds T_1 and T_2 are experimentally chosen as 0.45 and 0.50.

3.4.4.1 Performance measures

To evaluate the performance, we use the overall pixel accuracy (OP) [104] metric. The OP measures the proportion of correctly labeled pixels. We can compute OP as follows.

$$OP = \frac{\sum_{i=1}^{L} C_{ii}}{\sum_{i=1}^{L} G_i} \tag{3.20}$$

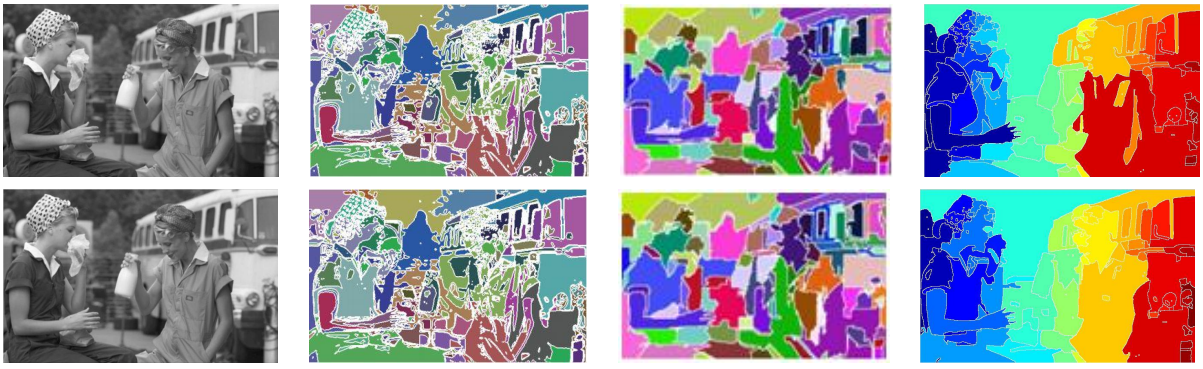
Where C is confusion matrix and $G_i = \sum_{j=1}^{L} C_{ij}$, is the total number of pixels labeled with i. L is number of classes.

3.4.4.2 Evaluation Dataset

Experiments are carried out over two different types of datasets, one acquired with a static camera (NYU depth dataset) [105] and the other acquired with a moving camera (NYU Scene Dataset) [87, 96].

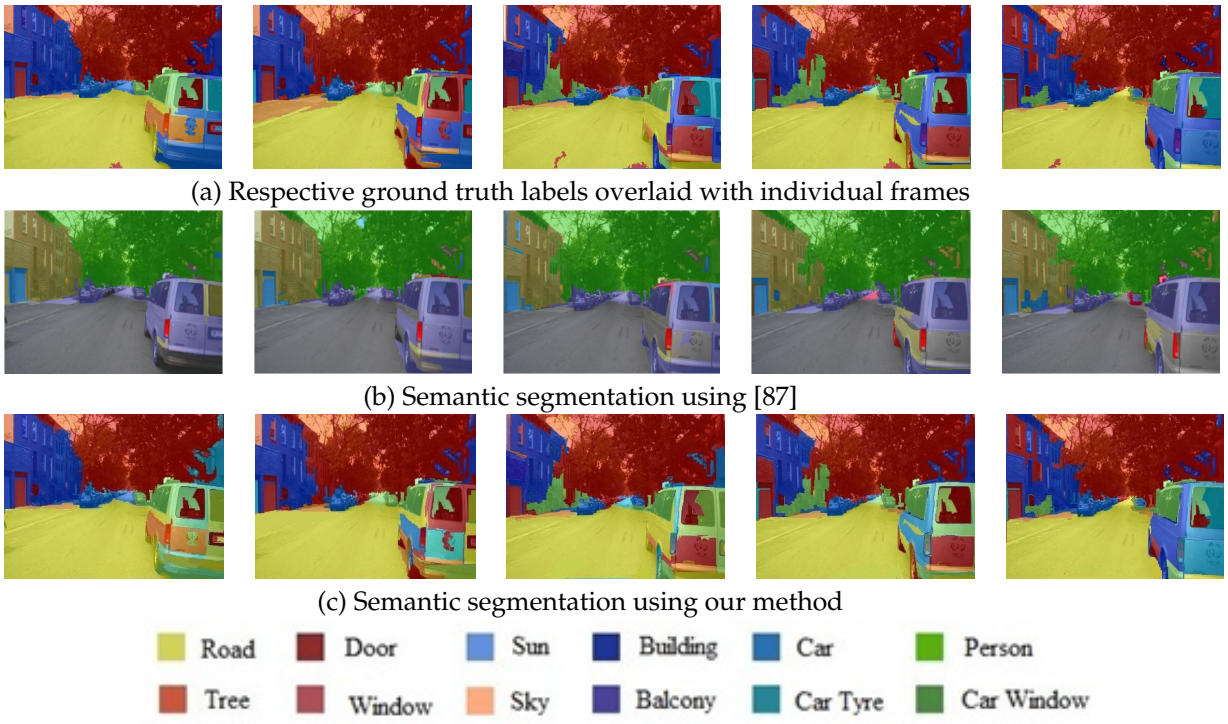
3.4.4.3 Performance comparison for causal video segmentation

To demonstrate the robustness of our method in terms of spatial consistency we compare our results with that of [87] and [97] in Fig. 3.4. For our experiment, we use 500 superpixels (an experimentally chosen value) for each frame. In case of the NYU scene dataset, the results are shown in Table 3.1. In this table, we compare our method with the results of frame by frame method, [96] and [87]. Table 3.1 clearly demonstrates the OP of our method (85.63) is superior as compared to that of the frame by frame (71.11), [96] (75.31), and [87] (76.27). We also show in Table 3.1 that the the modified DBSCAN (OP: 85.63) yield better results than the standard DBSCAN (OP: 78.26). In fig. 3.5, we present the comparison of our semantic segmentation with the ground truth and with that of [87] for five intermediate frames 55 - 59 of the NYU Scene dataset. The labeled images are overlaid on the original frames for better representation. The results clearly show that our output frames resemble the ground truth much better as compared to that of [87]. The quantitative results in terms of overall pixel accuracy (OP) for the NYU Depth dataset are presented in Table 3.2. We experiment with four videos from the NYU Depth dataset, namely, Dining room, Living room, Classroom and Office. Our proposed method (using modified DBSCAN) with an average OP of 72.32 surpasses both the frame-by-frame approach with an OP of 60.5 and that of [87] with an average OP of 61.6.



Original Frame | Mean Shift [97] | Couprie et al. [87] | Our Results

FIGURE 3.4: Comparison of spatially consistent segments on different frames of Two women dataset [97] with independent segmentation.



(a) Respective ground truth labels overlaid with individual frames

(b) Semantic segmentation using [87]

(c) Semantic segmentation using our method

FIGURE 3.5: Comparison of temporally consistent semantic video segmentation on frames 55 - 59 of NYU Scene dataset.

TABLE 3.1: OP values for the semantic segmentation task on the NYU Scene dataset.

	frame by frame	Miksik et al. [96]	Couprie et al. [87]	Proposed method	
				DBSCAN [99] for initial frame	*modified DBSCAN for initial frame*
Accuracy	71.11	75.31	76.27	**78.26**	**85.63**

TABLE 3.2: OP for the semantic segmentation task on the NYU Depth dataset.

Dataset	Frame by frame	Couprie et al. [87]	Proposed Method With Modified DBSCAN
Dining room	63.8	58.5	**78.80**
Living room	65.4	72.1	**83.28**
Classroom	56.5	58.3	**65.55**
Office	56.3	57.4	**61.63**
Mean :	60.5	61.6	**72.32**

3.4.5 Discussions

In this work as an initial step towards providing solution to multispectral video fusion, we present a solution for the problem of causal video segmentation using superseeds and local graph matching. The superseeds are selected from the superpixels extracted using the SLIC algorithm. The labels of the superseeds are propagated using local graph matching. Finally, watershed algorithm is used to obtain the complete segmentation. In future, we will work on improving the execution time of our method. We will also explore how the segmentation accuracy can be further improved.

3.5 Proposed Causal Multispectral Video Fusion

The multispectral videos to be fused are assumed to be registered in space and time. Our three-step solution pipeline begins with a two-part video pre-processing step. In the first part of the pre-processing, the superpixels are extracted from the current frame in both the spectra (VIS, IR/TH-IR (henceforth will be denoted IR)) using SLIC [89]. Next, we identify superpixels which can potentially be in motion. Superpixel level spatial and temporal saliency maps are obtained in the second step of the pipeline. In the third and final step, superpixels are categorized based on their saliency values and four different fusion rules are developed. Fig. 3.6 shows the proposed framework.

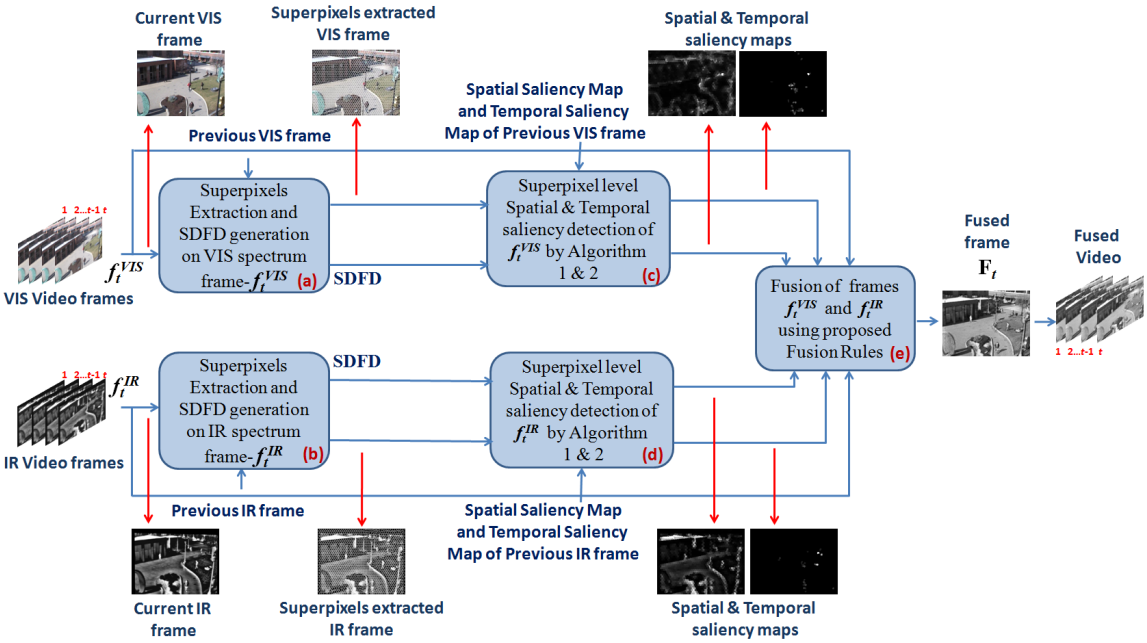


FIGURE 3.6: Framework: Proposed Causal multispectral Video Fusion.

3.5.1 Video pre-processing

The well-known SLIC [89] algorithm is used for the extraction of the superpixels in each frame of VIS and IR videos. Superpixel segmentation is given by:

$$S_t = SLIC(t, k) \tag{3.21}$$

Here, S_t denotes the superpixels extracted from the current frame t and k denotes the desired number of superpixels. To find the superpixels which can potentially be in motion, we employ superpixel level direct frame difference (SDFD) expressed as:

$$SDFD_{S_{i,t}} = g(S_{i,t}) - g(S_{i^*,t-1}) \tag{3.22}$$

Here, $g(S_{i,t})$ and $g(S_{i^*,t-1})$ respectively denote the mean intensities of $S_{i,t}$, the i^{th} superpixel in the current frame t and $S_{i^*,t-1}$, the co-located superpixel (i^*) in the previous frame $(t-1)$. The term $SDFD_{S_{i,t}}$ represents the difference in the two intensity values for the superpixel $S_{i,t}$ and is compared with an experimentally chosen threshold (T_1). The superpixels, for which $SDFD_{S_{i,t}}$ are larger than T_1, are deemed to be in motion (marked by motion labels $\phi_{S_{i,t}}$ with value 1). So, we can build a binary motion map:

$$\Phi_{S_{i,t}} = \begin{cases} 1 & \text{if} \quad SDFD_{S_{i,t}} \geq T_1 \\ 0 & \text{else} \end{cases} \tag{3.23}$$

The binary motion map for the frame f_t can be expressed as $\Phi_t = \cup_{i=1}^{n} \Phi_{S_{i,t}}$, where n is the actual number of superpixels in the frame f_t.

3.5.2 Saliency models for Video

Saliency detection in an image or video aims at extracting regions which capture greater attention of human vision system as compared to other portions [106]. There exist many pixel based saliency models in spatial [107, 108] as well as in frequency domains [13, 106]. Recently, Liu et al. [109] proposed a superpixel level spatio-temporal saliency detection algorithm. Although, this method outperforms several state-of-the-art saliency models in terms of accuracy, it suffers from low computational efficiency due to the use of pixel level optical flow algorithm for superpixel motion estimation. We propose here a computationally efficient superpixel based causal spatio-temporal saliency model.

3.5.2.1 Spatial saliency detection

We use superpixel level texture and color measures for spatial saliency. CIELAB space is used to obtain color measure due to its perceptual uniformity [89]. The relationship of a given superpixel $S_{i,t}$ with its first order neighborhood superpixels is explored to obtain its color measure $\Upsilon_{S_{i,t}}$. So, we write:

$$\Upsilon_{S_{i,t}} = \frac{1}{N}\sum_{j=1}^{N} \|\overline{Lab_{S_{i,t}}} - \overline{Lab_{S_{j,t}}}\|_2 \tag{3.24}$$

Here, $S_{j,t}$ is a first-order neighboring superpixel and N is the total number of first-order neighboring superpixels of $S_{i,t}$. The term $\overline{Lab_{S_{i,t}}}$ denotes the ***Lab*** vector with L representing luminance and a, b representing chroma components. Note that the IR spectrum contains only the luminance information.

The texture measure is based on the superpixel level local binary pattern (SLBP) [100]. The SLBP number characterizes the local structure and can be computed as follows:

$$SLBP_{P,R}(S_c) = \sum_{p=1}^{P} \theta(g(S_p) - g(S_c))2^p \tag{3.25}$$

where P denotes the number of superpixels within a circular neighborhood of radius R centering superpixel S_c. The terms $g(S_c)$ and $g(S_p)$ respectively represent mean intensities of the center superpixel S_c and a neighborhood superpixel S_p. We fix $R = 1$ for this work. The function θ is given by:

$$\theta(x) = \begin{cases} 1 & \text{if} \quad x \geq 0 \\ 0 & otherwise \end{cases} \tag{3.26}$$

We use $\Gamma_{S_{i,t}}$ to denote the texture measure of the superpixel $S_{i,t}$ and this is given by:

$$\Gamma_{S_{i,t}} = \frac{1}{N}\sum_{j=1}^{N} W_H\left(SLBP_{P,R}(S_{i,t}) \oplus SLBP_{P,R}(S_{j,t})\right) \tag{3.27}$$

W_H is the Hamming weight function on each SLBP and $\oplus$ stands for the logical XOR operation. The spatial saliency of the superpixel $S_{i,t}$ is expressed as:

$$\Psi_{S_{i,t}} = \Upsilon_{S_{i,t}} * \Gamma_{S_{i,t}} \tag{3.28}$$

The spatial saliency map for the frame f_t can be expressed as $\Psi_t = \cup_{i=1}^{n} \Psi_{S_{i,t}}$. We now present Algorithm 4 to obtain Ψ_t.

Algorithm 4 Spatial Saliency Detection

Input: S_t, ϕ_t, Ψ_{t-1}, n

Output: Ψ_t

Initialization: $\Psi_{S_{i,t}} \leftarrow \Psi_{S_{i^*,t-1}}, i = 1, \cdots, n$

1: **for** i = 1 to n **do**
2: **if** $(\phi_{S_{i,t}} == 1)$ **then**
3: Obtain $\Upsilon_{S_{i,t}}$
4: Obtain $\Gamma_{S_{i,t}}$
5: $\Psi_{S_{i,t}} \leftarrow \Upsilon_{S_{i,t}} * \Gamma_{S_{i,t}}$
6: **end if**
7: **end for**
8: $\Psi_t = \cup_{i=1}^{n} \Psi_{S_{i,t}}$

Please note that the spatial saliency map of the current frame ($\Psi_{S_{i,t}}$) is initialized with the spatial saliency map of the previous frame ($\Psi_{S_{i^*,t-1}}$). So, we essentially build the spatial saliency maps for different frames in a causal manner.

3.5.2.2 Temporal saliency detection

In our proposed model, we obtain temporal saliency for only those superpixels in the current frame which have non-zero motion. A temporal matching scheme is applied between these superpixels in the current frame with the co-located superpixels in the previous frame using local region graphs. A dissimilarity score based on the

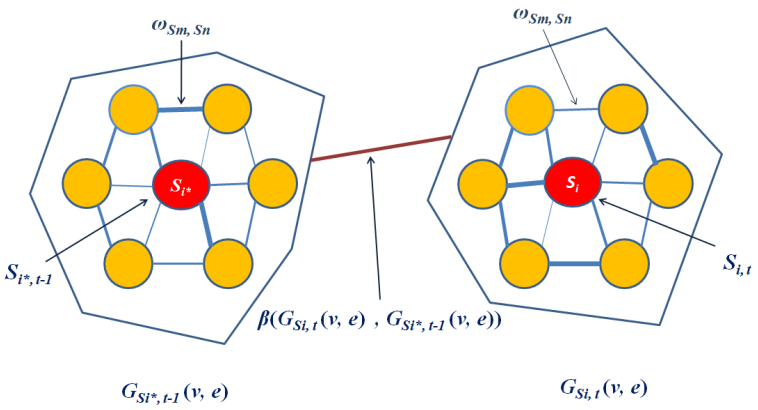


FIGURE 3.7: Local region graph matching.

eigenvalues of graph laplacians are employed for determining temporal saliency. The details of the proposed temporal saliency detection strategy are given below.

3.5.2.2.1 Local region graph construction: Local region graphs $G_{S_{i,t}}(v,e)$ and $G_{S_{i^*,t-1}}(v,e)$ are constructed surrounding the superpixels i in motion in the current frame t and at the co-located superpixels i^* in the previous frame $(t-1)$. Neighboring superpixels of $S_{i,t}$ and $S_{i^*,t-1}$ form the respective vertex sets. In each of the graphs $G_{S_{i,t}}(v,e)$ and $G_{S_{i^*,t-1}}(v,e)$, edges are now constructed between the center vertex and all neighbor vertices plus between each pair of neighbor vertices. The local region graphs are illustrated in figure 3.7. Spatial affinity between any two superpixels is obtained as a product of color and texture affinities between them. This value is assigned as the corresponding edge weight. The color affinity $C(S_m,S_n)$ and texture affinity $T(S_m,S_n)$ between superpixels S_m and S_n in a graph are given by:

$$C(S_m,S_n) = \|\overline{Lab_{S_m}} - \overline{Lab_{S_n}}\|_2 \tag{3.29}$$

$$T(S_m,S_n) = W_H\left(SLBP_{P,R}(S_m) \oplus SLBP_{P,R}(S_n)\right) \tag{3.30}$$

So, the edge weight between the two superpixels/vertices S_m and S_n is expressed as:

$$\omega_{S_m,S_n} = C(S_m,S_n) * T(S_m,S_n) \tag{3.31}$$

3.5.2.2.2 Local region graph matching [101]: We use the spectral graph theory based approach for the local region graph matching. The adjacency matrix of a weighted graph captures the edge weights. The degree matrix represents the number of edges connected to each vertex and is hence diagonal in nature. Let A_1, A_2 be the weighted adjacency matrices, D_1, D_2 be the diagonal degree matrices, and L_1, L_2 be the Laplacian matrices of the graphs $G_{S_{i,t}}(v,e)$ and $G_{S_{i^*,t-1}}(v,e)$ respectively. Then, the Laplacian matrices are given by:

$$\begin{aligned} L_1 &= D_1 - A_1 \\ L_2 &= D_2 - A_2 \end{aligned} \tag{3.32}$$

Dissimilarity score β between the two graphs $G_{S_{i,t}}(v,e)$ and $G_{S_{i^*,t-1}}(v,e)$ is given by the differences of top K eigenvalues $(\lambda_{11} \cdots \lambda_{1K})$ of L_1 and $(\lambda_{21} \cdots \lambda_{2K})$ of L_2. So, we can write:

$$\beta(G_{S_{i,t}}(v,e), G_{S_{i^*,t-1}}(v,e)) = \sum_{k=1}^{K} (\lambda_{1k} - \lambda_{2k})^2 \tag{3.33}$$

Top K eigenvalues are the ones which contain 90% of the energy. Hence, K is determined using the following equation:

$$\min_{q \in [1,2], p} \left(\frac{\sum_{p=1}^{K} \lambda_{qp}}{\sum_{p=1}^{M} \lambda_{qp}} > 0.9 \right) \tag{3.34}$$

In the above equation, M represents the total number of eigenvalues. A high value of β indicates that the graphs are highly dissimilar. The local graph matching is graphically illustrated in Fig. 3.7.

3.5.2.2.3 Building temporal saliency map: The dissimilarity value β between the two region graphs centering two co-located superpixels is compared with an experimentally chosen threshold T_2. If the dissimilarity value is less than T_2, then co-located superpixel (S_{i^*})'s temporal saliency value from the previous frame $(t-1)$ is deemed

as the temporal saliency value of the superpixel (S_i) in the current frame t. Otherwise, the dissimilarity value itself is assigned as the saliency value of the superpixel under consideration. The temporal saliency map for the frame f_t can be expressed as $\Omega_t = \cup_{i=1}^{n} \Omega_{S_{i,t}}$. We now present Algorithm 5 to obtain Ω_t.

Algorithm 5 Temporal Saliency Detection

Input: S_t, ϕ_t, Ω_{t-1}, n

Output: Ω_t

Initialization: $\Omega_{S_{i,t}} \leftarrow 0, i = 1 \cdots n$

for i = 1 to n **do**
  if $(\phi_{S_{i,t}} == 1)$ **then**
    Build $G_{S_{i,t}}(v, e)$
    Build $G_{S_{i^*,t-1}}(v, e)$
    Obtain $\beta(G_{S_{i,t}}(v, e), G_{S_{i^*,t-1}}(v, e))$
    if $\beta(G_{S_{i,t}}(v, e), G_{S_{i^*,t-1}}(v, e)) < T_2$ **then**
      $\Omega_{S_{i,t}} \leftarrow \Omega_{S_{i^*,t-1}}$
    else
      $\Omega_{S_{i,t}} \leftarrow \beta(G_{S_{i,t}}(v, e), G_{S_{i^*,t-1}}(v, e))$
    end if
  end if
end for
$\Omega_t = \cup_{i=1}^{n} \Omega_{S_{i,t}}$

Please note that to derive the temporal saliency map of the current frame, our approach makes use of temporal saliency map of previous frame. Temporal saliency values of the superpixels in motion in the current frame are initialized with the co-located superpixel's temporal saliency values from the previous frame. So, we also develop the temporal saliency maps in a causal manner.

3.5.3 Rules for video fusion

Depending on the spatial and temporal saliency values of the current frame in the VIS and IR spectra, we divide the superpixels in the two spectra into four groups, namely, *Uniform*, *Spatially salient*, *Temporally salient*, and *Spatio-temporally salient*. For the sake of brevity, we omit the subscript t from the symbols denoting superpixels, spatial saliency map and temporal saliency map as by default only the current frame will be referred. On a similar note, we add the subscripts *VIS* and *IR* to denote these quantities in two different spectra. We now discuss below how the superpixels are categorized into four groups.

1. *Uniform Superpixels*: In both the spectra, the temporal saliency values of the superpixels are zero and the spatial saliency values are below the mean.

2. *Spatially salient Superpixels*: In both the spectra, the temporal saliency values of the superpixels are zero and in at least one spectrum the spatial saliency value is above the mean.

3. *Temporally salient Superpixels*: In at least one spectrum, the temporal saliency value of the superpixels is non-zero and in both the spectra the spatial saliency values are below the mean.

4. *Spatio-temporally salient Superpixels*: In at least one spectrum, the temporal saliency value of the superpixels is non-zero and in at least one spectrum the spatial saliency value is above the mean.

Separate fusion rules are applied for superpixels in the above four categories to obtain the finally fused video. These fusion rules are now described below.

3.5.3.1 Fusion rule for uniform superpixels

The superpixels classified as uniform represent homogeneous background regions in a scene. In many transform domain and spatial domain fusion methods weighted average fusion rule is adopted to fuse such regions. Accordingly, we propose the following superpixel level energy modulated fusion rule:

$$F(S_i^j) = \frac{E_{S_{i,VIS}}}{E_{S_{i,VIS}} + E_{S_{i,IR}}} * f(S_{i,VIS}^j) + \frac{E_{S_{i,IR}}}{E_{S_{i,VIS}} + E_{S_{i,IR}}} * f(S_{i,IR}^j) \tag{3.35}$$

Here, $F(S_i^j)$ represents the intensity of the j^{th} pixel in the i^{th} superpixel S_i in the current fused frame. The terms $f(S_{i,VIS}^j)$ and $f(S_{i,IR}^j)$ denote the intensities of the same pixel in VIS and IR spectrum respectively. Similarly, $E_{S_{i,VIS}}$ is the normalized energy of the superpixel S_i in VIS spectrum and $E_{S_{i,IR}}$ is the normalized energy of the same superpixel in the IR spectrum. The expression for the normalized energy is given by:

$$E_{S_i} = \frac{(g(S_i))^2}{\max_i [g(S_i))^2]} \tag{3.36}$$

Where, $g(S_i)$ is the mean intensity of the superpixel S_i.

3.5.3.2 Fusion rule for spatially salient superpixels

When a superpixel is identified as spatially salient, we use the spatial saliency value in the fusion process. Normalized energy is used in conjunction to ensure robustness. Energy modulated spatial saliency fusion rule is thus formulated in the following

manner:

$$F(S_i^j) = \frac{(E_{S_{i,VIS}} + \Psi_{S_{i,VIS}})}{(E_{S_{i,T}} + \Psi_{S_{i,T}})} * f(S_{i,VIS}^j) + \frac{(E_{S_{i,IR}} + \Psi_{S_{i,IR}})}{(E_{S_{i,T}} + \Psi_{S_{i,T}})} * f(S_{i,IR}^j) \tag{3.37}$$

Where,

$$E_{S_{i,T}} = \frac{E_{S_{i,VIS}} + E_{S_{i,IR}}}{\max_i [E_{S_{i,VIS}} + E_{S_{i,IR}}]}, \qquad \Psi_{S_{i,T}} = \frac{\Psi_{S_{i,VIS}} + \Psi_{S_{i,IR}}}{\max_i [\Psi_{S_{i,VIS}} + \Psi_{S_{i,IR}}]}.$$

Here, $\Psi_{S_{i,VIS}}$ and $\Psi_{S_{i,IR}}$ are the spatial saliency values of the superpixel S_i in the current VIS and IR spectrum frame. The terms $E_{S_{i,T}}$ and $\Psi_{S_{i,T}}$ respectively denote total energy and total spatial saliency of a pixel by taking into consideration both the spectra.

3.5.3.3 Fusion rule for temporally salient superpixels

When a superpixel is identified as temporally salient, we make use of the temporal saliency value in the fusion process. Normalized energy is used in conjunction, as in previous cases, to ensure robustness. Energy modulated temporal saliency fusion rule is thus formulated in the following manner:

$$F(S_i^j) = \frac{(E_{S_{i,VIS}} + \Omega_{S_{i,VIS}})}{(E_{S_{i,T}} + \Omega_{S_{i,T}})} * f(S_{i,VIS}^j) + \frac{(E_{S_{i,IR}} + \Omega_{S_{i,IR}})}{(E_{S_{i,T}} + \Omega_{S_{i,T}})} * f(S_{i,IR}^j) \tag{3.38}$$

Where,

$$\Omega_{S_{i,T}} = \frac{\Omega_{S_{i,VIS}} + \Omega_{S_{i,IR}}}{\max_i [\Omega_{S_{i,VIS}} + \Omega_{S_{i,IR}}]}.$$

Here, $\Omega_{S_{i,VIS}}$ and $\Omega_{S_{i,IR}}$ denote the temporal saliency values of the superpixel S_i in the current VIS and IR spectrum frame respectively. The term $\Omega_{S_{i,T}}$ represents total temporal saliency of a pixel from both the spectra.

Algorithm 6 Video fusion

Input: $\Psi_{S_i,VIS}$, $\Psi_{S_i,IR}$, $\Omega_{S_i,VIS}$, $\Omega_{S_i,IR}$, n.

Output: F: Fused frame.

for i = 1 to n **do**
  Obtain $\Psi^{'}_{VIS}$, mean spatial saliency in VIS spectrum
  Obtain $\Psi^{'}_{IR}$, mean spatial saliency in IR spectrum
  if $((\Omega_{S_{i,VIS}} == 0)\ \ \&\&\ \ (\Omega_{S_{i,IR}} == 0))$ **then**
    if $((\Psi_{S_{i,VIS}} < \Psi^{'}_{VIS})\ \ \&\&\ \ (\Psi_{S_{i,IR}} < \Psi^{'}_{IR}))$ **then**
      Apply fusion rule 1 using eqn. 3.35
    else
      Apply fusion rule 2 using eqn. 3.37
    end if
  else
    if $((\Psi_{S_{i,VIS}} < \Psi^{'}_{VIS})\ \ \&\&\ \ (\Psi_{S_{i,IR}} < \Psi^{'}_{IR}))$ **then**
      Apply fusion rule 3 using eqn. 3.38
    else
      Apply fusion rule 4 using eqn. 3.39
    end if
  end if
end for
Return F

3.5.3.4 Fusion rule for spatio-temporally salient superpixels

Finally, we frame fusion rules for spatio-temporally salient superpixels. Since a superpixel in this case is both spatially and temporally salient in nature, we obtain the spatio-temporal values in each spectrum by combining the spatial and temporal saliency values. Then, we employ energy modulated spatio-temporal saliency values as a part of the fusion rule in the following manner:

$$F(S_i^j) = \frac{(E_{S_{i,VIS}} + \Lambda_{S_{i,VIS}})}{(E_{S_{i,T}} + \Lambda_{S_{i,T}})} * f(S_{i,VIS}^j) + \frac{(E_{S_{i,IR}} + \Lambda_{S_{i,IR}})}{(E_{S_{i,T}} + \Lambda_{S_{i,T}})} * f(S_{i,IR}^j) \tag{3.39}$$

Where,

$$\Lambda_{S_{i,VIS}} = \frac{\Psi_{S_{i,VIS}} + \Omega_{S_{i,VIS}}}{\max_i[\Psi_{S_{i,VIS}} + \Omega_{S_{i,VIS}}]}$$

$$\Lambda_{S_{i,IR}} = \frac{\Psi_{S_{i,IR}} + \Omega_{S_{i,IR}}}{\max_i[\Psi_{S_{i,IR}} + \Omega_{S_{i,IR}}]}$$

$$\Lambda_{S_{i,T}} = \frac{\Lambda_{S_{i,VIS}} + \Lambda_{S_{i,IR}}}{\max_i[\Lambda_{S_{i,VIS}} + \Lambda_{S_{i,IR}}]}$$

Here, $\Lambda_{S_{i,VIS}}$ and $\Lambda_{S_{i,IR}}$ are the spatio-temporal saliency values of the superpixel S_i in VIS and IR spectrum. The term $\Lambda_{S_{i,T}}$ represents total spatio-temporal saliency of a pixel from both the spectra. We now present algorithm 6 to obtain the fused video frame F.

3.5.4 Time-Complexity Analysis

The complexity of the pre-processing step to extract superpixels is $O(N)$ [89], where N is the number of pixels in a video frame. Let n be the number of superpixels. Then, superpixel level SDFD generation is done in $O(n)$. The complexity of spatial saliency detection requires color and texture measures. The color measure can be obtained in $O(kn)$, where k is the maximum number of the first order neighboring superpixels. The texture measure is based on superpixel level LBP (SLBP) which has $O(n^2)$ complexity [110]. So, the total complexity of spatial saliency detection is $O(n^2) + O(kn) = O(n^2)$ as $(k << n)$. The complexity of temporal saliency detection involves local region graph construction and matching. Let m be the number of superpixels in motion. Complexity of constructing m Laplacian matrices is $O(mk^2)$ and complexity of obtaining their eigenvalues is $O(mk^3)$. So, the complexity of the temporal saliency detection is $O(mk^3)$. The fusion rules on the superpixels can be applied in linear time with a complexity of $O(n)$. So, the frame level time-complexity of our pipeline is $O(n) + O(n^2) + O(kn) + O(mk^2) + O(mk^3) + O(n)$ = $O(n^2)$.

3.5.5 Experimental results

3.5.5.1 Evaluation Dataset, Comparisons and Performance measures

Experiments are carried out on five publicly available multispectral (VIS and IR) video dataset, namely, Video pairs 1-5. Video pair 1 consists of pair (VIS and IR) of 1651 frames, whereas video pair 2 and 3 consists of 750 and 600 pairs of frames respectively [111]. All these videos are acquired by fixed outdoor thermal Sensor (Raytheon PalmIR 250D, 25 mm lens) and Color Sensor (Sony TRV87 Handycam) with 240×320 pixels resolution and sampling rate of 30 Hz. Video pair 4 with 8702 pair of frames is acquired during night time and contains natural noise [112] with 240×320 pixels resolution and sampling rate of 25 Hz. Video pair 5, popularly known as Bristol Eden Project Multi-Sensor Data Set 3 [113] is acquired by moving cameras with 480×576 pixels resolution and sampling rate of 25 Hz. It consists of 100 registered VIS and IR spectrum frames.

We also include experiments done with four more publicly available multispectral video datasets (Video pair 6,7,8 and 9) [114] containing 250 to 2300 frames of size 658×491. The aquisition of these videos is performed using commercial camera from FluxData Inc. (the FD-1665-MS) with varying frame rates which depends on overall scene illumination, i.e., 5 frames/sec (for dark one) to 15 frames/sec (for bright one). Please note that we could only show comparisons with the alternate saliency model of [109] on these datasets. This is because the results of other fusion methods for these datasets are not available. All the experiments are performed on a PC with Intel Core $i7$ processor having $3.4\ GHz$ speed and $8\ GB$ RAM.

The proposed CMVF algorithm is compared with recent multisensor multispectral video fusion methods like ST-Maximum, ST-Matching, ST-Liang-HOSVD, ST-PCNN, ST-Structure-Tensor and ST-HOSVD1 as reported in [14]. Furthermore, comparisons are also included with methods based on DWT, DT-CWT, 3D-DWT, 3D-DTCWT, 3D-UDCT-salience, MCWT as can be found in [25]. We also compare our work with a

recently reported video fusion algorithm for real-time target detection [86]. To show the robustness of our framework and to justify our choice of SLIC for superpixels extraction, we also compare our SLIC superpixels based framework (CMVF) with that of Lazy Random Walk (LRW) [115] superpixels based framework (LRW-CMVF). In addition, we evaluate our work with a recently proposed superpixel based spatio-temporal saliency model [109]. For the objective evaluation of video fusion performance we use five metrics- IE(Information entropy), Q_m, DQ_G, IFD_MI and t_c (per frame execution time in sec.). Higher value of IE is desirable. The measure Q_m [116] signifies how much of the salient information contained in each of the input videos has been transferred into the fused video without introducing any distortions or artifacts. Q_m varies in the range [0, 1] where a value of 1 corresponds to the best fusion performance. Dynamic fusion quality index, DQ_G [117] is an extension of gradient information preservation between the input and the fused frames. It is based on preservation of spatial information estimates obtained from the current frame and temporal information preservation estimates obtained from the previous and subsequent frames. The metric $DQ^{AB/f}$ also refers to the same Dynamic fusion quality index metric proposed in [117]. The dynamic range of DQ_G [25] or $DQ^{AB/f}$ [14] is [0, 1] and high value signifies better performance. IFD_MI [3] metric is used to denote the temporal stability of a fused video. A temporally stable and consistent video fusion method is marked by a high value of IFD_MI.

3.5.5.2 Selection of k, T_1, T_2

There are three parameters in our proposed pipeline, which we have selected experimentally. The first parameter is the number of superpixels, k, used as an input to SLIC. In Fig. 3.8, we show the variations of the sum of the averages of the three objective measures (IE, Q_m, $DQ_G/DQ^{AB/f}$) with k for video pairs $1-5$. The curves saturate after 1700 superpixels for the video pairs $1, 2, 3, 4$ with a 240×320 resolution. However, for the video pair 5 with a higher resolution of 480×576, the same

curve saturate at 2200. So, for video pairs 1-4 we set $k = 1700$, while for video pair 5 we set $k = 2200$. The second parameter T_1 is used to threshold the SDFD map to determine candidate superpixels which can be in motion. The third parameter is the threshold T_2 used for obtaining the temporal saliency map. In Fig. 3.9, we show a surface plot showing variations of the sum of the averages of the same three objective measures (IE, Q_m, $DQ_G/DQ^{AB/f}$) with T_1 and T_2. From the figure, we find that the best performance is obtained with T_2=0.10 and T_1=0.15.

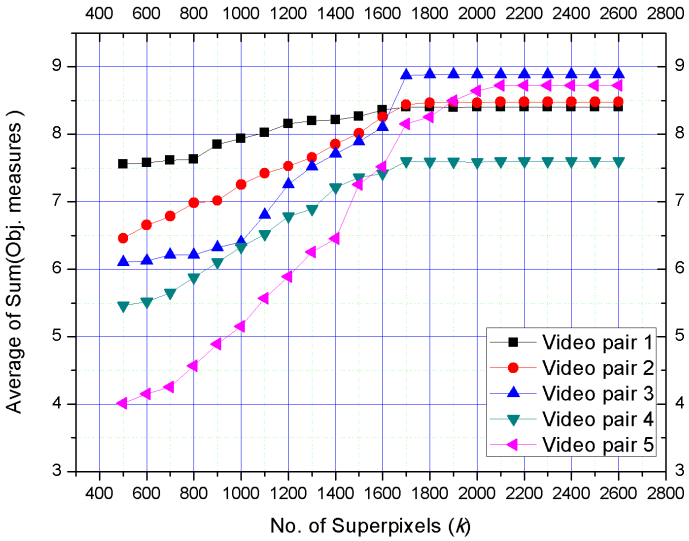


FIGURE 3.8: Estimation of number of superpixels, k.

3.5.5.3 Effectiveness of fusion rules

We now demonstrate the effectiveness of the four fusion rules using Fig. 3.10. Here the experiments are performed by turning on only one fusion rule at a time for all the superpixels. The average values of the four objective measures (IE, Q_m, $DQ_G/DQ^{AB/f}$, IFD_{MI}) are shown with the independent activation of fusion rules and the full combination. The curves clearly demonstrate IE and IFD_{MI} improves significantly when all four fusion rules are fired. For, $DQ_G/DQ^{AB/f}$ and IFD_{MI}, the improvement is marginal. So, overall, we can surely say that firing of all four fusion rules are necessary to improve the performance.

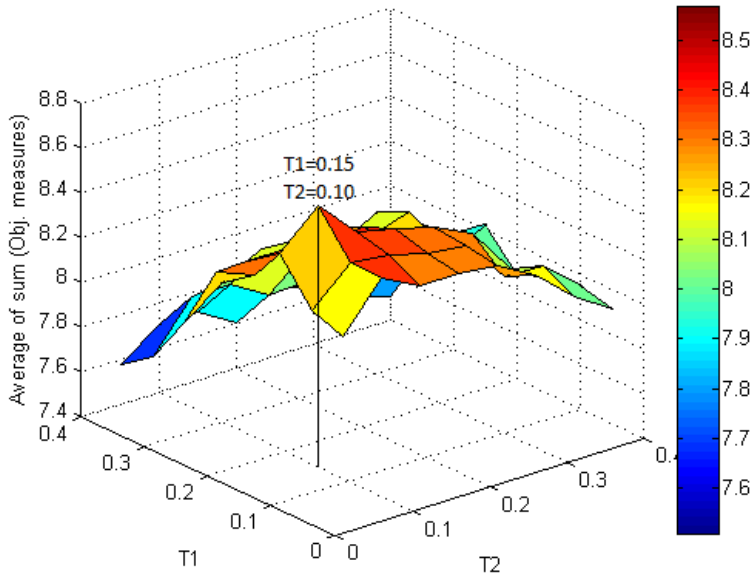


FIGURE 3.9: Estimation for threshold values: T_1 and T_2.

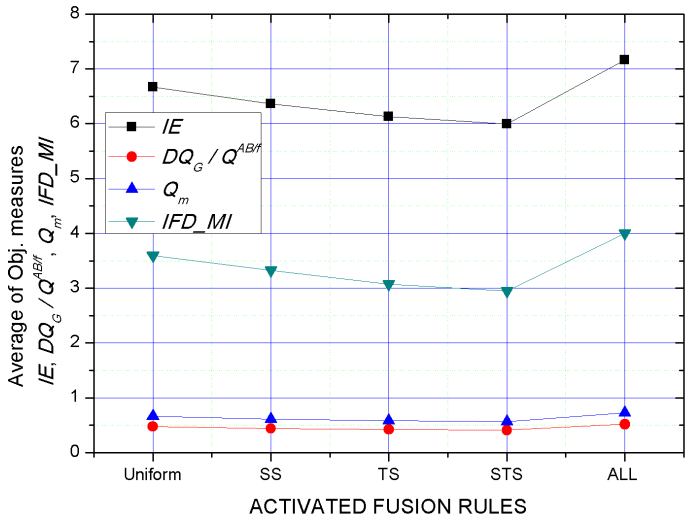


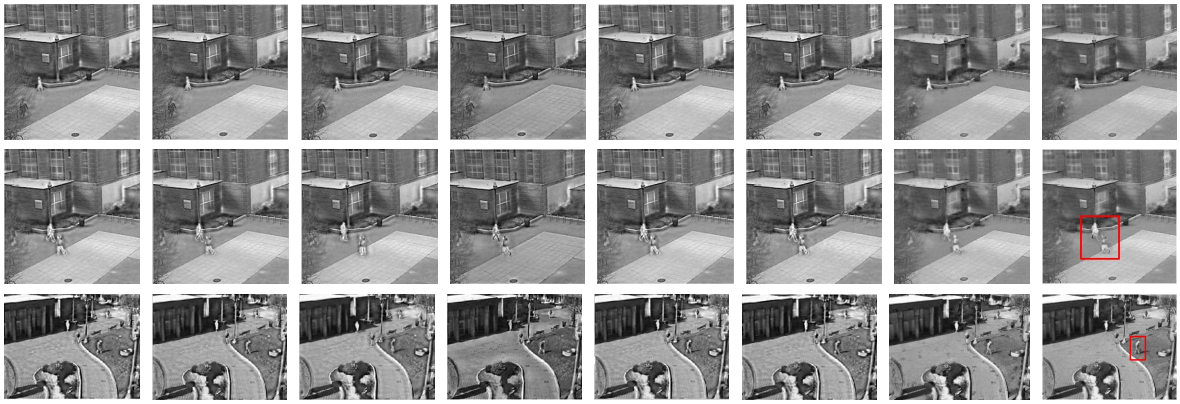
FIGURE 3.10: Improvement achieved by combination of the proposed fusion rules. Here Uniform, SS, TS, STS and ALL represents fusion rules for Uniform, Spatially salient, Temporally salient, Spatio-temporally salient superpixels and all fusion rules respectively.

3.5.5.4 Performance comparison for causal video fusion

Comparison with other video fusion methods:

We first provide a qualitative comparison among the different competing methods. Fig. 3.12 shows fused frames by the CMVF, LRW-CMVF, and some transform domain

FIGURE 3.11: Sample video pair frames (VIS and IR): (a, f) Video pair 1, (b, g) Video pair 2, (c, h) Video pair 3, (d, i) Video pair 4, (e, j) Video pair 5 (EDEN).



ST-Maximum ST-Matching ST-Liang ST-PCNN ST-Struct-Tensor ST-HOSVD1 LRW-CMVF CMVF

FIGURE 3.12: Fused frames obtained using different methods [14]: Row 1- Fused frame number 634 from Video Pair 1, Row 2- Fused frame number 98 from Video Pair 2. Row 3- Fused frame number 242 from Video Pair 3.

methods reported in [14]. The results of our algorithm compare well on Video pair 1 and are certainly better on Video pairs 2 and 3. This is highlighted by using the red rectangles over certain regions in the fused frame. The moving object (human) in the highlighted regions suffer from halo effect (bordering) in case of the transform domain methods as compared to the proposed method. Furthermore, the minute details in the background of the object are also missing for the other methods. From the visual inspection, it is clear that the quality of fused frames from CMVF and LRW-CMVF are comparable.The improvement in our method stems from the use of precise spatial and temporal saliency detection and novel region based fusion rules. Fig. 3.14 shows the intermediate outputs at different stages of our proposed fusion pipeline on six adjacent frames from Video pair 3. We next provide fused results on

three consecutive frames from Video pair 4 (frames- 1096-1-1099) and Video pair 5 (frames- 50-1-53) in Fig. 3.15. Qualitatively the fused results for Video pair 4 show the robustness of the proposed method against natural noise. Fig. 3.15 also illustrates the background details for the fused frames in Video pair 5 are not affected by the dynamic background produced due to explicit ego motion. The foreground moving object is also clearly visible without any artifacts like halo effects.

Table 3.3 shows quantitative comparisons of CMVF, LRW-CMVF and the methods reported in [14] on Video pairs 1, 2 and 3. The performance of the proposed method is comparable in terms of IE. In terms of Q_m and DQ_G our method surpasses others by a considerable margin. The improvement in Q_m is a clear indication of very less inconsistencies and instabilities in the fused video as obtained using our method. Improvement in DQ_G validates the high accuracy of transferring the spatio-temporal gradient information from input videos to the fused one. The performance of LRW-CMVF and CMVF are quite comparable in terms of the performance metrics (except for superpixel extraction times). See Table 3.4 for quantitative comparison with other transform domain methods reported in [25] on Video pairs 3, 4 and 5. Once again, the IE values are quite comparable. There is consistent improvement in $DQ^{AB/f}$ and IFD_MI which corroborate the robustness of our method. In Table 3.5, we provide quantitative comparison with a recently reported frame-by-frame real-time video fusion algorithm for target detection application [86]. The experimental results are available only on Video pair 5. In terms of the evaluation metrics $Q_{AB/f}$ and IE the results are very comparable.

Comparison with an alternative saliency model:

As stated earlier, Liu et al. [109] have recently proposed a superpixel based spatio-temporal saliency model. We first derive the spatio-temporal saliency maps using the codes made available by the authors. Then, we use these spatio-temporal maps for generating the fused frames. In Fig. 3.16, we have shown fused frames obtained using the proposed saliency model and that of from the saliency model described

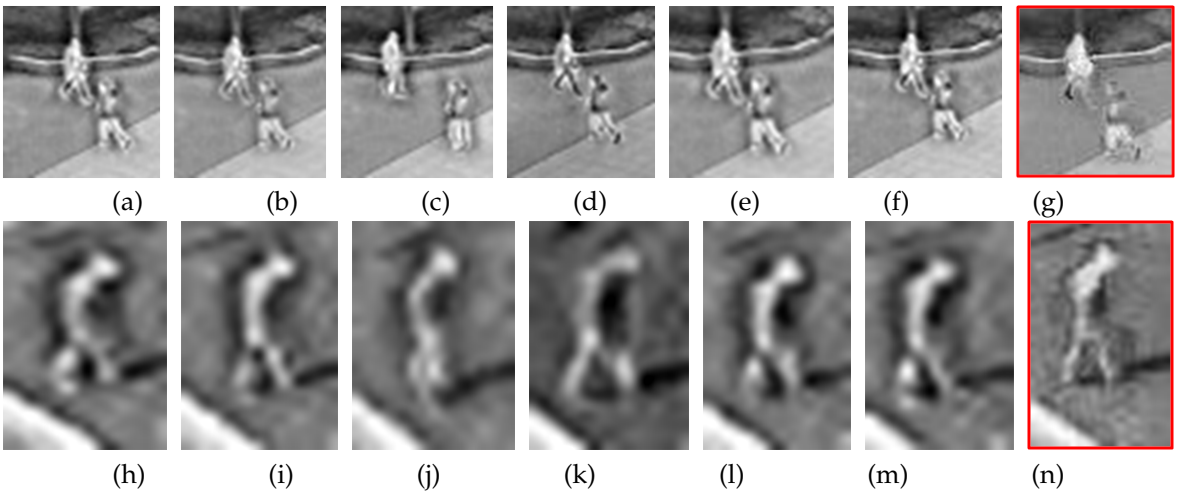


FIGURE 3.13: Magnified part of fused frame from Video pair 2 (row 1) and 3 (row 2), see Fig. 3.12: (a, h) ST-Maximum, (b, i) ST-Matching, (c, j) ST-Liang-HOSVD, (d, k) ST-PCNN, (e, l) ST-struct-tensor, (f, m) ST-HOSVD1, (g, n) Proposed method.

TABLE 3.3: Video fusion quantitative results: Performance comparison of CMVF, LRW-CMVF with results reported in [14] on Video pairs 1, 2 and 3.

Dataset	Metric	ST-Maximum	ST-Matching	ST-Liang -HOSVD	ST-PCNN	ST-Structure -Tensor	ST-HOSVD1	LRW -CMVF	CMVF
Video pair 1	IE	7.3847	7.3585	7.3809	7.2045	7.3662	**7.3848**	7.3330	7.2705
	Q_m	0.6184	0.6235	0.6170	0.6262	0.6291	0.6290	0.6804	**0.6984**
	DQ_G	0.2974	0.3032	0.2969	0.3048	0.3059	0.3060	0.4262	**0.4360**
	t_c (sec.)	2.9549	5.278	211.1204	41.9252	49.1361	2.4700	2.0410	**2.0410**
Video pair 2	IE	7.4078	7.3792	7.3951	7.2143	7.3880	**7.4128**	7.3395	7.3131
	Q_m	0.5973	0.6032	0.5951	0.6061	0.6094	0.6092	0.6717	**0.6920**
	DQ_G	0.2827	0.2861	0.2813	0.2978	0.2934	0.2908	0.4255	**0.4363**
	t_c (sec.)	2.9547	5.2979	212.1545	42.0132	47.1739	2.4809	2.0226	**2.0226**
Video pair 3	IE	7.7046	7.6811	7.6994	7.6024	7.6841	**7.7065**	7.3123	7.5936
	Q_m	0.6802	0.6854	0.6775	0.6975	0.6915	0.6885	0.7331	**0.7566**
	DQ_G	0.3795	0.3828	0.3713	0.3751	0.3918	0.3901	0.5053	**0.5228**
	t_c (sec.)	3.6386	6.5679	267.3234	52.1245	60.7098	3.0264	2.0906	**2.0906**

TABLE 3.4: Video fusion quantitative results: Performance comparison of CMVF, LRW-CMVF with results reported in [25] on video pairs 3, 4 and 5.

Dataset	Metric	DWT	DT-CWT	3D-DWT	3D-DTCWT	3D-UDCT salience	MCWT	LRW-CMVF	CMVF
Video pair 3	IE	7.6746	7.6958	7.6785	7.6867	7.6965	**7.8409**	7.3123	7.5936
	$DQ^{AB/f}$	0.2274	0.2643	0.2894	0.3012	0.3105	0.3052	0.5053	**0.5228**
	IFD_MI	0.5729	0.7986	1.0598	1.2045	1.2064	1.2451	3.9339	**3.5409**
Video pair 4	IE	6.4051	6.2674	6.3344	6.2491	6.3896	**6.4784**	6.1651	6.2441
	$DQ^{AB/f}$	0.3354	0.3377	0.3599	0.3753	0.3902	0.3907	0.6186	**0.6289**
	IFD_MI	1.6349	2.0251	2.9869	2.9556	2.5314	2.2133	**4.4755**	4.3577
Video pair 5	IE	7.0215	7.027	7.1132	6.8326	7.0479	7.0733	6.8612	**7.4024**
	$DQ^{AB/f}$	0.4027	0.385	0.4004	0.3313	0.4373	0.4413	0.4373	**0.5641**
	IFD_MI	2.7754	2.8942	3.0353	3.4662	3.3514	3.5356	**5.1041**	5.0610

TABLE 3.5: Video Fusion Quantitative results: Comparison with [86] on Video pair 5.

Frame	$Q_{AB/f}$					**Entropy**					t_c **(sec.)**				
	[118]	[119]	[120]	[86]	CMVF	[118]	[119]	[120]	[86]	CMVF	[118]	[119]	[120]	[86]	CMVF
25	0.4316	0.5257	0.3766	0.4624	0.4550	6.7691	7.1236	0.3766	7.3158	7.0018	36.5	542	224	38	**5.3962**
50	0.4300	0.5247	0.3791	0.4609	0.4533	6.7331	7.0700	7.0064	7.2600	6.9668					
50	0.4449	0.5229	0.4041	0.4336	0.4231	6.8509	6.8509	6.8782	7.1844	6.8292					
50	0.456	0.5192	0.4039	0.4419	0.4344	6.7156	6.7156	6.8660	7.2114	6.8347					

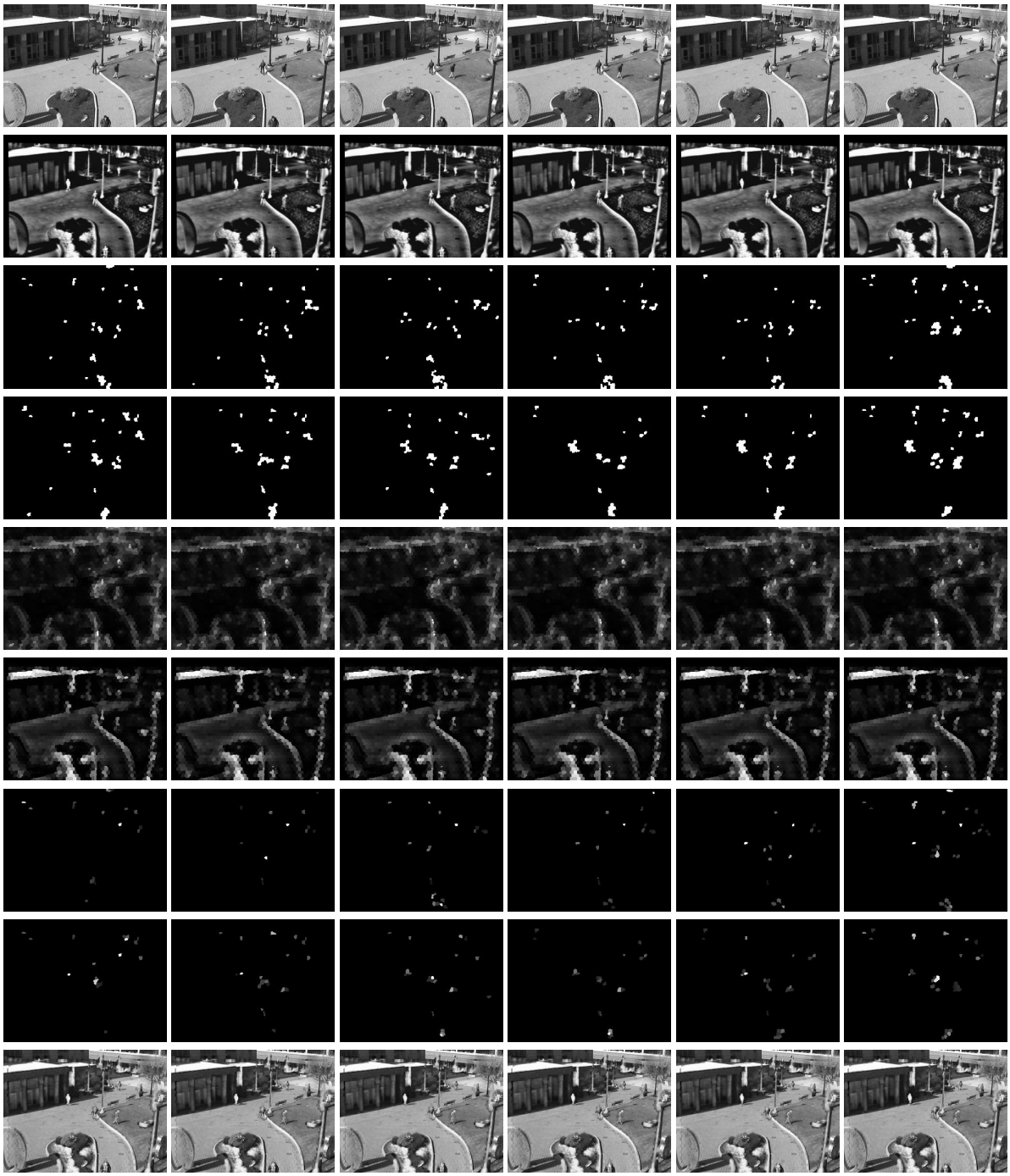


FIGURE 3.14: Output at various intermediate stages of the proposed method (CMVF). Row 1: VIS frames 112-2-122, Row 2: IR frames 113-2-123, Row 3: SDFD maps of VIS frames 112-2-122, Row 4: SDFD maps of IR frames 113-2-123, Row 5: Ψ_t of VIS frames 112-2-122, Row 6: Ψ_t of IR frames 113-2-123, Row 7: Ω_t of VIS frames 112-2-122, Row 8: Ω_t of IR frames 113-2-123, Row 9: Corresponding fused frames.

in [109]. The regions highlighted within green rectangles in the Fig. 3.16 clearly indicate that the integration of complementary information from the VIS and the IR spectrum is more accurate for our model, especially in the temporally salient regions.

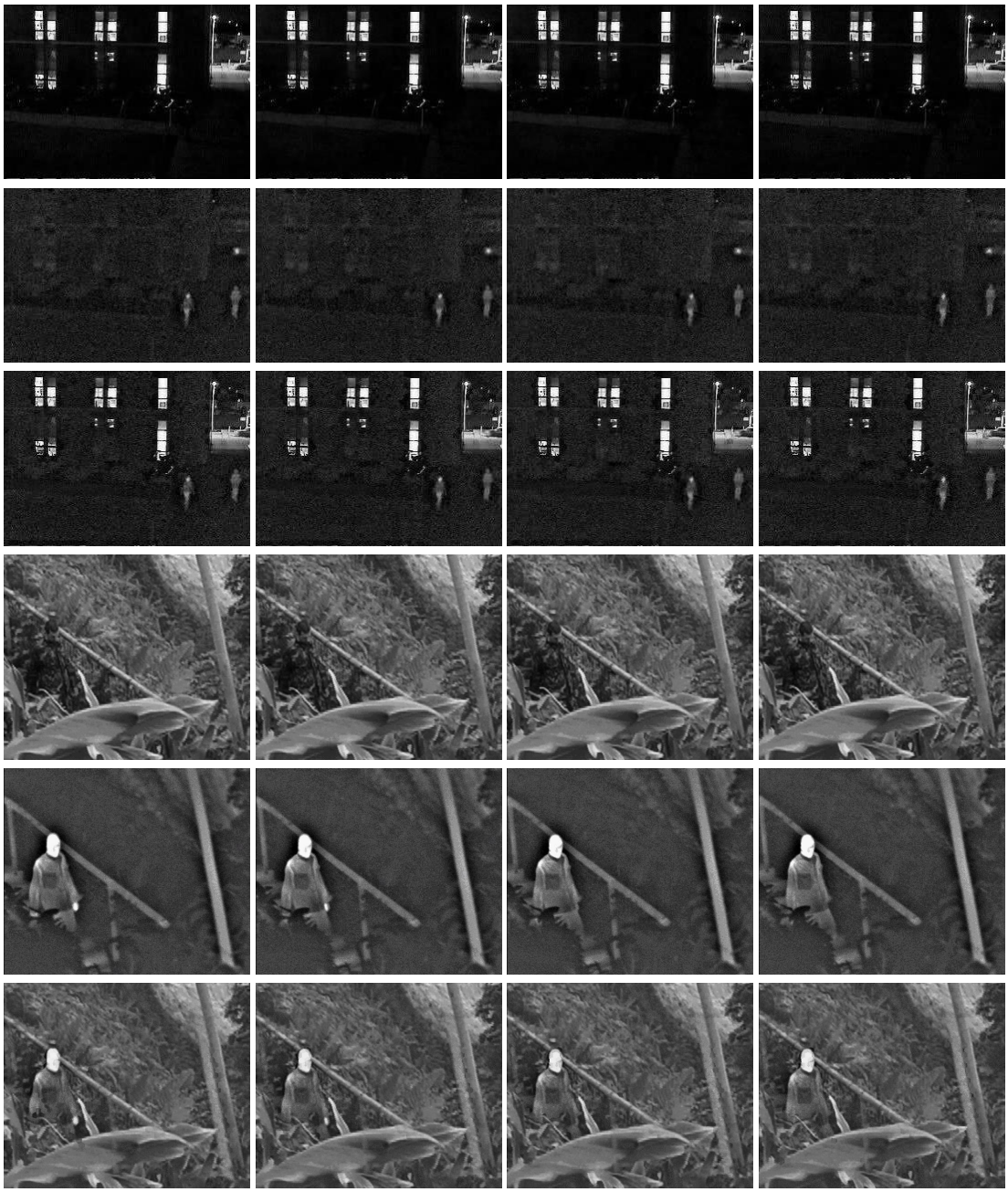


FIGURE 3.15: Fused results on Video pair 4 and 5. Row 1: VIS frames of Video pair 4, Row 2: IR frames of Video pair 4, Row 3: Fused frames of Video pair 4, Row 4: VIS frames of Video pair 5, Row 5: IR frames of Video pair 5, Row 6: Fused frames of Video pair 5.

In addition, the background regions highlighted in yellow rectangles demonstrate that Liu et al.'s model introduces some artifacts at these locations. We also compare the two approaches quantitatively on all nine video pairs in Table 3.6. The values of the three performance metrics, IE, Q_m, DQ_G and t_c clearly indicate the superiority of the proposed saliency model which resulted in better quality fused frames.

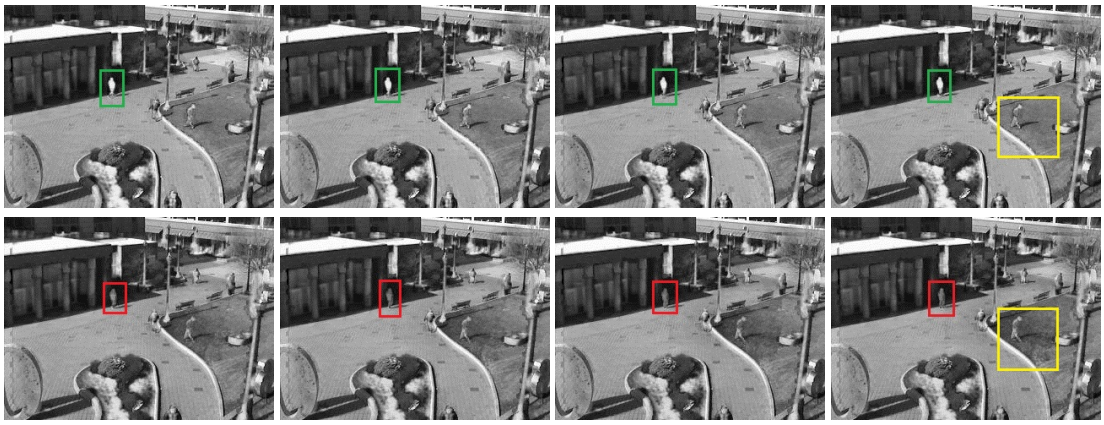


FIGURE 3.16: Comparison between Liu et al. [109] saliency model and proposed saliency model. Row 1: Fused frames 116-2-122 using proposed saliency model, Row 2: Fused frames 116-2-122 using Liu et al.'s saliency model.

TABLE 3.6: Video fusion performance comparison with the use Liu et al.'s [109] spatio-temporal saliency model.

Method	Dataset	IE	Q_m	DQ_G	t_c (sec.)
[109]	Video pair 1	7.2703	0.6810	0.4080	5.9543
	Video pair 2	7.2973	0.6756	0.4250	6.2407
	Video pair 3	7.5939	0.7510	0.5105	5.5701
	Video pair 4	6.0821	0.6957	0.6508	18.6034
	Video pair 5	7.2280	0.7154	0.5279	18.4500
	Video pair 6	7.4074	0.8669	0.6603	16.2362
	Video pair 7	7.5643	0.8642	0.5709	16.3452
	Video pair 8	7.1458	0.8360	0.5536	16.8762
	Video pair 9	7.4192	0.7972	0.5402	15.9543
	Average	6.8980	0.7507	0.5866	16.3529
CMVF	Video pair 1	7.2705	0.6984	0.4360	2.0410
	Video pair 2	7.3131	0. 6920	0.4363	2.0226
	Video pair 3	7.5936	0.7566	0.5228	2.0906
	Video pair 4	6.3365	0.7241	0.6805	14.2080
	Video pair 5	7.3802	0.7581	0.5797	14.3410
	Video pair 6	7.4101	0.8763	0.6718	12.5351
	Video pair 7	7.6014	0.8838	0.6101	12.5248
	Video pair 8	7.1529	0.8546	0.5740	12.5211
	Video pair 9	7.4201	0.8026	0.5458	12.5301
	Average	**7.0376**	**0.7788**	**0.6416**	**13.6892**

Comparison of execution times:

We exclude the SLIC superpixel extraction time from the total execution time of the proposed solution pipeline. It is imperative that a faster superpixel segmentation algorithm would further improve the overall execution time of our method. First, we show using Table 3.3 our method outperforms all other transform domain methods including the recently proposed [14] in terms of the average execution time. On

an average, our method takes only 2.0514 sec. per frame as compared to 2.6595 sec. by ST-HOSVD1 [14], ST-Maximum (3.1827 sec.), ST-Matching (5.7146 sec.), ST-Lian-HOSVD (230.1994 sec.), ST-PCNN (45.3543 sec.), ST-Structure-Tensor (52.3399 sec.). For the video pair 5, the algorithm in [86] takes about 38 sec. per frame to process. This is superior to NSCT [120] (224 sec.) and HOSVD [119] (542 sec.) and is comparable to the Laplacian method [118] which consumes 36.5 sec. As shown in Table 3.5, our method takes only 5.3962 sec. on average to process each frame of the same dataset. Use of superpixel level processing results in great computational saving over that of the pixel level. For example, time consumed per frame by pixel level processing for various stages of our framework are: 2.0400 sec. (spatial saliency detection), 37.4700 sec. (temporal saliency detection) and 0.1577 sec. (fusion) resulting in a total time of 39.6677 sec. In contrast, the total time of superpixel level processing is only 2.0514 sec. We also demonstrate in Table 3.6 that fusion performed with the proposed superpixel level saliency model is faster (13.6892 sec. vs. 16.3529 sec., excluding SLIC superpixel extraction time in both the methods) as compared to fusion based on the alternative superpixel level saliency model of [109]. From the implementations made available by the respective authors, we found that LRW based superpixel extraction (226.245 sec./frame on average) is approximately 32 times slower than SLIC (7.0283 sec./frame on average) for the Video pairs 1-4; and is approximately 68 times slower than SLIC (1162.7600 sec./frame vs. 16.9451 sec./frame on average) for the high resolution Video pair 5. This also validates the use of SLIC in our framework.

3.5.6 Discussions

In this work, we proposed a superpixel level causal multispectral video fusion algorithm. Visible and infrared video pairs are fused using this algorithm to obtain highly accurate information in a time-efficient manner. Comprehensive comparison with several existing approaches on a number of publicly available datasets clearly indicate the advantage of our fusion method. In future, we will examine if superpixel extraction can be made faster which in turn would further reduce the execution time of the proposed algorithm. Another direction of future research will be to analyze the fused video to solve important surveillance tasks like anomalous event detection [8] and person re-identification [9].

Chapter 4

Multimodal biometric system using 2D and 3D Palmprints

Multi-biometric recognition systems have become very popular as a counter measure for direct spoofing attack. Success of such systems depend heavily on designing effective fusion schemes, which can combine complementary information from multiple traits. Palmprint has evolved as a popular trait over the years due to non-intrusiveness, low cost for capture device, and stable structure features. But 2D and 3D palmprint in isolation can become vulnerable to spoofing attacks. In this chapter we propose a multimodal biometric system based on a novel combination of 2D and 3D palmprints. We first generate 2.5D palmprint data using standard deviation based signal level fusion of 2D and 3D palmprints. A graph based template matching framework is designed for the purpose of recognition. Comprehensive comparisons with several existing works indicate the benefit of our solution.

4.1 Introduction

The most convenient and reliable way of identification or verification of persons is generally based on some physiological or behavioral attributes of the individuals. These characteristics like face, fingerprint, iris, palmprint, ear, gait, voice, retina are commonly referred as *biometrics* or *traits* or *cues* [20]. Biometric systems are widely being used for several security based applications. In recent years, multi-biometric approaches, which use complementary information from multiple traits, have become very popular as a counter measure for spoofing [15]. These information can be

integrated at different levels and we can subdivide them into two main categories- prior to matching/pre-classification fusion and after matching/post-classification fusion. In prior to matching/pre-classification case, information fusion takes place before matching and can be obtained at signal level (low level) and at feature level (mid level). In case of after-matching/post-classification, information fusion takes place after matching at score level, rank level and decision level fusion (high level) [16–18, 30]. Biometric systems that integrate information at an early stage of processing are believed to be more effective than those systems which perform integration at a later stage [19]. Feature level fusion suffers from unknown relationship between feature spaces of different modalities and curse of dimensionality. Fusion at decision level is often too rigid since only limited amount of information is available at this level. Overall, the level at which fusion is performed plays a crucial role in the robustness of the system.

Palmprint has evolved as a popular biometric trait due to its non-intrusiveness, low cost for capture device, and stable features. For works on 2D palmprint based recognition, one can see [121–131]. Majority of these works used PolyU palmprint database with a resolution of 100 PPI [132]. At such a low resolution, ridges and valleys cannot be observed and matching is mainly based on texture information in form of principal lines (flexion creases), and major and minor wrinkles (secondary creases). However, 2D palmprint is not very robust to illumination changes and contamination on palms and can be quite vulnerable to spoofing attacks[26]. The human palmprint is not plain but essentially three dimensional (3D) in nature. Utilizing the 3D palmprint information can improve the recognition performance. Range images acquired by 3D sensor contain palmprint surface shape information [27]. 3D palmprint is preferred over other 3D biometric traits like 3D face [133–135] and 3D ear [28, 29]. 3D palmprint is not affected by various facial expressions and is much easier to acquire and more user-friendly than 3D ear. To build more robust and highly

secure palmprint recognition system we can take advantage of the highly discriminative texture rich information from 2D palmprint and depth information from 3D palmprint. Only a few works have been reported on fusion of the 2D and the 3D palmprint information [26, 31–33]. Majority of these approaches are based either on feature or score level fusion. To the best of our knowledge, there is no work reported on signal/sensor level fusion of 2D and 3D palmprint data.

In this work we propose a multimodal biometric system based on fusion of 2D and 3D palmprints. First, we fuse the aligned regions of interest (ROI) of 2D and 3D palmprints to produce 2.5D palmprint data. The coarse texture information from 2D and the fine texture and depth information from 3D are integrated in the 2.5D data using standard deviation based fusion rule. The fused 2.5D palmprint representation is proved to be more informative than the original individual 2D and 3D palmprints. In the second stage, we use graph based template matching framework to derive the matching scores between the test sample and the gallery samples.

4.2 Related works

We first start discussing 2D palmprint based methods. 2D palmprint based recognition techniques are classified into four categories: line based [124, 128, 129], subspace based [122, 125], Statistics based [121, 126, 127, 131] and Coding based [123, 130]. In [124], Sun et al. proposed a novel palmprint representation method based on orthogonal line ordinal features. Huang et al. [128] proposed a novel palmprint verification approach based on principal lines and Radon transform. In [129], Jia et al. proposed a novel robust line orientation code for palmprint verification to achieve higher recognition rate and faster processing speed. Wu et al. in [122] proposed a novel method called Fisherpalms based on linear projection using Fisher's linear discriminant. In [125], Connie et al. tested and compared various linear subspace projection techniques like PCA, FDA and ICA. In [121] Li et al. proposed palmprint recognition in

frequency domain using Fourier transform. Shang et al. [126] proposed radial basis probabilistic neural network (RBPNN) based palmprint recognition system. In [131], Wei et al. proposed new descriptor of palmprint named histogram of oriented lines (HOL) to overcome certain disadvantages of subspace based systems like- high sensitivity to the illumination, translation, and rotation variances in image recognition. In [130], Feng et al. proposed Hashing based fast palmprint identification for large-scale databases. Very detailed survey of various 2D palmprint recognition systems and use of fusion with other traits can be found in [136].

Relatively less works are reported on 3D palmprint based recognition [31, 137, 138]. Majority of these works have explored MCI (Mean Curvature Image) and GCI (Gaussian Curvature Image) for matching and classification. [31] Zhang et al. exploited 3D structural depth information from the range data of palmprint. In [137], Ni et al. proposed a novel 3D palmprint recognition using Dempster-Shafer fusion theory. Same GCI and MCI based features are extracted and used for the purpose. The fusion of these features is proposed by belief function determined by the Dempster-Shafer (D-S) fusion theory. In [138], Li et al. presents a very simple and efficient scheme for 3D palmprint recognition using the line and orientation features extracted from the enhanced mean curvature image (MCI).

We now discuss some works on fusion of 2D and 3D palmprint data. In [26], Zhang et al. proposed robust and accurate multilevel 2D and 3D palmprint based authentication system. They used the surface curvature based 3D features and Gabor feature based competitive coding scheme for representing 2D features. They have shown that match score level fusion causes the improvement in recognition performance as compared to that of 2D and 3D features in isolation. In [32], Li et al. used features at shape level, line level and texture level. Shape level features are taken from the 3D palmprint, line level features and fine texture features are collected from both the 2D and 3D palmprints. The texture information is used for palmprint discrimination and the shape and line features are used for refining and alignment purpose. A

novel matching scheme is proposed to efficiently use features at three different levels for accurate palmprint verification. In [33], Meraoumia et al. have proposed an efficient multi-biometric system based on 2D and 3D palmprint. Rotation invariant variance based features are extracted and compressed using PCA from both modalities. Further, the feature vector of each palmprint is modeled by Hidden Markov Model (HMM). Finally, the log-likelihood matching score level fusion is used to integrate the individual scores of 2D and 3D palmprint recognition.

To the best of our knowledge, there is no work on signal/sensor level fusion of 2D and 3D palmprint data reported in the literature. The two main contributions of this work are now highlighted below:

1. Signal level fusion scheme to fuse 2D and 3D palmprints to produce more infromative 2.5D palmprint, and

2. Novel Graph based template generation and matching scheme on 2.5D palmprint data

4.3 Proposed method

Fig. 4.1 shows the block diagram of our multi-biometric recognition system. The three major components of the proposed system are: (i) Guided filtering based enhancement (pre-processing), (ii) Signal level fusion of 2D and 3D palmprint data and (iii) Graph based recognition. We now discuss these major components in greater details.

4.3.1 Guided Filter based Enhancement

Guided filter [139, 140] is an edge-preserving smoothing filter, derived from a local linear model between guidance I and output q. The guided filter computes the

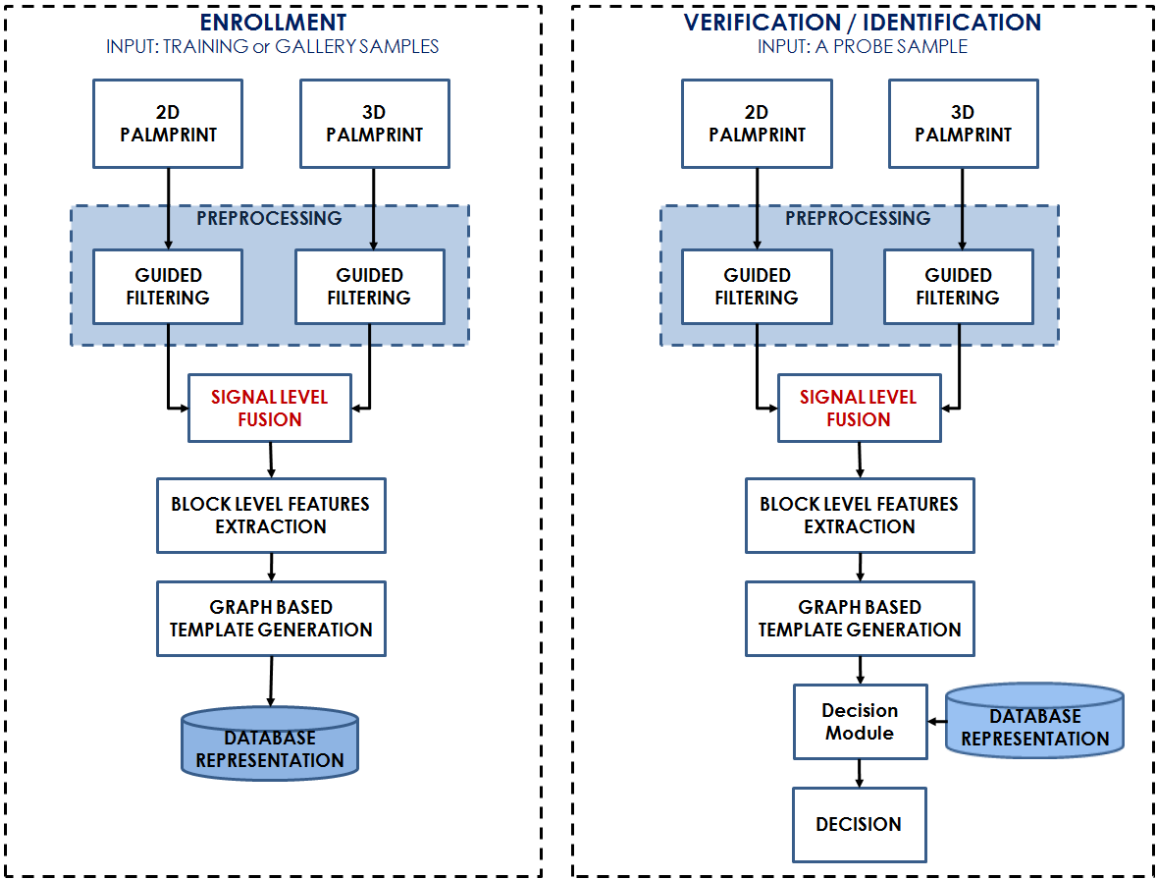


FIGURE 4.1: Schematic diagram of the proposed method

output by taking into account the content of the guidance image which can be the filtering input itself [70, 71]. Assuming q as a linear transform of I in window ω_k centered at pixel k, q can be expressed for a pixel i as shown below.

$$q_i = a_k I_i + b_k \quad \forall \ i \in \omega_k \tag{4.1}$$

Here, (a_k, b_k) are linear coefficients assumed to be constant in the window ω_k. As $\bigtriangledown(q) = a. \bigtriangledown (I)$, the guided filter preserves edges [140]. The linear coefficients are determined using constraints between filtering input p and q given as:

$$q_i = p_i - n_i \tag{4.2}$$

Where, n_i refers to unwanted noise components. We seek a solution that minimizes the difference between q(output) and p(input) while maintaining the linear model equation 4.1. Specifically, we minimize the following cost function in the window ω_k:

$$E(a_k, b_k) = \sum_{i \in \omega_k} ((a_k I_i + b_k - p_i)^2 + \epsilon a_k^2) \tag{4.3}$$

Where ϵ is regularization parameter penalizing large a_k. Above equation is linear regression model and its solution is given by.

$$a_k = \frac{\frac{1}{|\omega|} \sum_{i \in \omega_k} I_i p_i - \mu_k \bar{p_k}}{\sigma_k^2 + \in} \tag{4.4}$$

$$b_k = \bar{p_k} - a_k \mu_k \tag{4.5}$$

Here, μ_k represents the mean of I in window ω_k and σ_k^2 represents the variance of the image in the same window. $|\omega|$ is the number of pixels in the window ω_k and $\bar{p_k}$ is the mean of p computed in that window. As a pixel i can be involved in overlapping windows, the final output should be obtained by taking the average of all possible values of q.

$$q_i = \bar{a_i} I_i + \bar{b_i} \tag{4.6}$$

Where, $\bar{a_i}$ and $\bar{b_i}$ are the average value of a_k and b_k of all windows overlapping i. We employ guided filtering for overall enhancement of the aligned regions of interest (ROI) of 2D and 3D palmprints using the following equations:

$$I_{2D}^E = (I_{2D} - I_{2D}^{GF}) * 5 + I_{2D}^{GF} \tag{4.7}$$

$$I_{3D}^E = (I_{3D} - I_{3D}^{GF}) * 5 + I_{3D}^{GF} \tag{4.8}$$

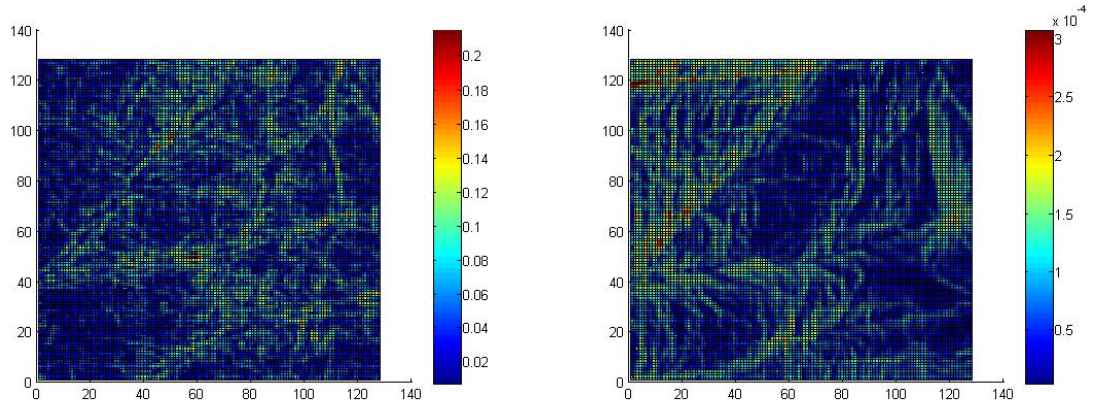
Here I_{2D} and I_{3D} are sample 2D and 3D palmprint images; I_{2D}^{GF} and I_{2D}^{GF} are guided filtered 2D and 3D palmprint images with guidance as same images and; I_{2D}^E and I_{3D}^E are enhanced 2D and 3D palmprints. The constant 5 used in the above equations is experimentally derived.

4.3.2 Signal level fusion of enhanced 2D and 3D palmprints

The signal or pixel level fusion is always preferred over feature and score level and other fusion techniques [19]. In signal level fusion, the pixel in fused image is produced by weighted linear combination of corresponding pixels in source images to be fused. The pixel level local standard deviation (SD) is a measure of local variance in intensity (in case of 2D palmprint) or depth (in case of 3D palmprint). The local variance is a measure of the local textural information. Hence, more the local SD, more will be the weight assigned while fusion of corresponding pixels from 2D and 3D palmprints. In particular, we set the weights as function of corresponding normalized local SD in the neighborhood of pixels to be fused. Thus, we adaptively fuse enhanced 2D (I^E_{2D}) and 3D (I^E_{3D}) palmprints to produce 2.5 D palmprint data. The coarse texture information from 2D and fine texture information and depth information from 3D are integrated in the fused 2.5D palmprint data. The adaptive fusion achieved is as follows.

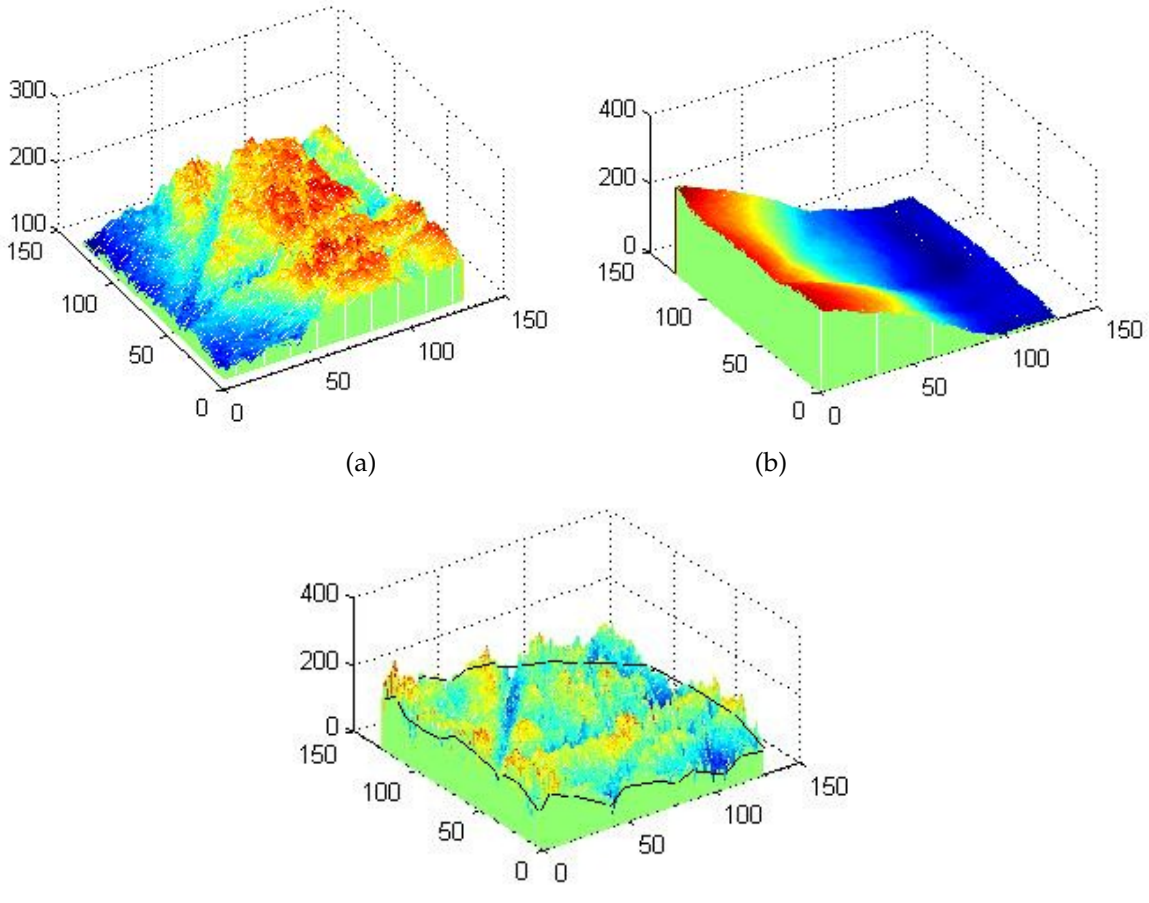
$$I_{2.5D}(i,j) = \begin{cases} \frac{\sigma_{2D}(i,j)}{\sigma_{2D}(i,j)+\sigma_{3D}(i,j)} * I^E_{2D}(i,j) + \\ \frac{\sigma_{3D}(i,j)}{\sigma_{2D}(i,j)+\sigma_{3D}(i,j)} * I^E_{3D}(i,j) \end{cases} \tag{4.9}$$

Here $\sigma_{2D}(i,j)$ is the weight determined by computing local standard deviation considering all the first order 8-neighborhood pixels of the pixel at location (i,j) in case of 2D palmprint. Similarly, we determine $\sigma_{3D}(i,j)$ considering all the first order 8-neighborhood pixels for a pixel at location (i,j) in case of 3D palmprint. The adaptive nature of weights used for fusing 2D and 3D enhanced palmprints is illustrated in Fig. 4.2. The fused 2.5D palmprint image is a result of successful integration of complementary information contained in 2D and 3D data. This is graphically illustrated by using mesh plots in Fig. 4.3. Fig. 4.3 (c) shows the fused 2.5D palmprint with approximately marked palmprint depth profile modulated by coarser texture information from 2D.



(a) (b)

FIGURE 4.2: Illustration: Adaptive nature of coefficient selection based on local standard deviation. (a) Coefficients/weights map adapted for 2D palmprint, (b) Coefficients/weights map adapted for 3D palmprint. (Subject 8-Sample 1)



(a) (b)

(c)

FIGURE 4.3: Illustration: Information integration into 2.5D palmprint shown using mesh plots. (a) 2D palmprint, (b) 3D palmprint, (c) 2.5D fused palmprint.

4.3.3 Graph based Matching

We propose a novel graph based template generation on each 2.5D sample. Graphs are very commonly used for the abstract representation of complex data with robust performance. The necessary details are described below.

4.3.3.1 Graph Construction

First, we divide 128×128 pixels of 2.5D palmprint image into uniform blocks of sizes 8×8. From each of these blocks 2D texture Haralick features and 3D surface type features are extracted. An undirected weighted graph with four connected neighborhood is constructed from 2.5D data. Each 8×8 block is deemed as a vertex of the graph. The differences in the feature values between adjacent vertices are assigned as the corresponding edge weights. As we show in the next subsection, each feature vector contains 13 Haralick features and 1 surface primitive based on Mean and Gaussian curvatures. This type of graph construction makes the processing quite fast as we only handle graphs with relatively small sizes (256 vertices in the current block based model vs. 16384 vertices if pixels were used as graph vertices). Furthermore, we can also make use of aggregate behavior of all 64 pixels in a block to extract more robust features.

4.3.3.2 Block based Feature extraction

Haralick textures [141] is well-known statistical method for quantifying textures and gives information about the image region such as homogenity, contrast, boundaries, and complexity. This approach has been widely applied in the biomedical imaging analysis [142] and is quite successful. Haralick texture features are computed from the Gray Level Co-occurrence Matrix (GLCM) of an image. The GLCM of given image is a square matrix of dimension $N_g \times N_g$, where N_g is the number of gray levels in the image. Element (i, j) in GLCM is computed by counting the number of

times pixel with value i is adjacent to a pixel with value j and dividing entire matrix by the total number of such comparisons made. Each entry of GLCM is therefore considered to be the probability that a pixel with value i will be found adjacent to a pixel of value j. In our framework, the input is individual blocks of uniformly segmented 2.5D So, the GLCM of a given block can be written as:

$$G = \begin{bmatrix} p(1,1) & p(1,2) & . & . & . & p(1,N_g) \\ p(2,1) & p(2,2) & . & . & . & p(2,N_g) \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ p(1,1) & p(N_g,2) & . & . & . & p(N_g,N_g) \end{bmatrix} \tag{4.10}$$

Here, adjacency can be defined in either of the four directions: horizontal, vertical, left diagonal and right diagonal. Thirteen Haralick's texture features statistics are then computed from each of these directional GLCMs. By averaging these statistics over four different directional GLCMs, rotation invariance is ensured. Let $p(i,j)$ is the $(i,j)^{th}$ element in G. Then, we can have:

- Angular second moment (ASM):

$$ASM = \sum_i \sum_j p(i,j)^2 \tag{4.11}$$

- Contrast (C):

$$C = \sum_{n=0}^{N_g-1} n^2 \left(\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) \right) \tag{4.12}$$

  Where, $|i-j| = n$.

- Correlation (CR):

$$CR = \frac{\sum_i \sum_j p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y} \tag{4.13}$$

Where, μ_x, μ_y, σ_x and σ_y are means and standard deviations of p_x and p_y, the partial probability functions on $p(i, j)$.

- Variance (V):

$$V = \sum_i \sum_j (i - \mu)^2 p(i, j) \tag{4.14}$$

- Inverse Difference Moment (IDM):

$$IDM = \sum_i \sum_j \frac{1}{1 + (i - j)^2} p(i, j) \tag{4.15}$$

- Sum Average (SA):

$$SA = \sum_{i=2}^{2N_g} i P_{x+y}(i) \tag{4.16}$$

Where x and y are the coordinates (row and column) of an entry in the GLCM and $P_{x+y}(i)$ is the probability of GLCM coordinates summing $x + y$.

- Sum Variance (SV):

$$SV = \sum_{i=2}^{2N_g} (i - SE)^2 P_{x+y}(i) \tag{4.17}$$

- Sum Entropy (SE):

$$SE = -\sum_{i=2}^{2N_g} P_{x+y}(i) log(P_{x+y}(i)) \tag{4.18}$$

- Entropy (E):

$$E = -\sum_i \sum_j p(i, j) log(p(i, j)) \tag{4.19}$$

- Difference Entropy(DE):

$$DE = -\sum_{i=0}^{N_g - 1} P_{x-y}(i) log(P_{x-y}(i)) \tag{4.20}$$

- Difference Variance (DV)

$$DV = -\sum_{i=0}^{N_g-1} i^2 P_{x-y}(i) \tag{4.21}$$

- Information measure of Correlation 1 ($IFC1$)

$$IFC1 = \frac{HXY - HXY1}{max(HX, HY)} \tag{4.22}$$

- Information measure of Correlation 2 ($IFC2$)

$$IFC2 = (1 - e^{[-2(HXY2-HXY)]^{\frac{1}{2}}}) \tag{4.23}$$

Where, $HXY = -\sum_i \sum_j p(i,j) log(p(i,j))$, HX, HY are the entropies of P_x and P_y.

$HXY1 = -\sum_i \sum_j p(i,j) log P_x(i) P_y(i)$,

$HXY2 = -\sum_i \sum_j P_x(i) P_y(i) log P_x(i) P_y(i)$

We also extract 3D depth features from 2.5D palmprint. Each point on the depth map can be classified into one of the eight surface primitive type (ST) [143]. Let the 3D surface of 2.5D palmprint is represented by $I_{2.5D}(i, j, f(i, j))$. The mean curvature image (MCI) H; and the Gaussian curvatures image (GCI) K of the 2.5D palmprint can be computed as follows [144].

$$H = \frac{(1 + f_x^2) f_{yy} + (1 + f_y^2) f_{xx} - 2 f_x f_y f_{xy}}{2(1 + f_x^2 + f_y^2)^{\frac{3}{2}}} \tag{4.24}$$

$$K = \frac{f_{xx} f_{yy} - f_{xy}^2}{(1 + f_x^2 + f_y^2)^2} \tag{4.25}$$

where f_x, f_y are the first-order and f_{xx}, f_{yy} and f_{xy} are the second-order partial

derivatives. There are eight fundamental viewpoint independent surface types (STs) that can be characterized using only the sign of the mean curvature (H) and Gaussian curvature (K) [143]. So, total nine STs can be defined as listed in Table 4.1. These include eight fundamental STs and one special case for $H = 0$ and $K > 0$. Each point in the 2.5D plamprint can be classified into one of the nine STs and is accordingly labeled from 1 to 9. The surface type (ST) of each pixel in 8×8 pixel block is determined using the above procedure. Then, we define the ST of an individual block as the ST of maximum number of its constituent pixels. Finally, we combine both Haralick's feature vector (13 features) and ST features (1 feature) to form a 14-dimensional feature vector for each block.

TABLE 4.1: Surface types Labels defined by signs of surface curvatures.

	$K > 0$	$K = 0$	$K < 0$
$H < 0$	Peak(ST=1)	Ridge(ST=2)	Saddle Ridge(ST=3)
$H = 0$	None (ST=4)	Flat(ST=5)	Minimal Surface(ST=6)
$H > 0$	Pit(ST=7)	Valley(ST=8)	Saddle Valley(ST=9)

4.3.3.3 Template matching

During verification, we compare the query template with the gallery templates by comparing the corresponding graphs. There exist several measures of graph similarity [101]. In the present work, we use Frobenius norm of the difference of the two adjacency matrices, a simple yet accurate measure, to determine how similar two graphs are. Let $G1_{2.5D}$ and $G2_{2.5D}$ are the graph templates of two samples to be compared with $A_{G1_{2.5D}}$ and $A_{G2_{2.5D}}$ as their corresponding adjacency matrices. The graph similarity score can then be written as:

$$\beta(G1_{2.5D}, G2_{2.5D}) = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} |A_{G1_{2.5D}}(i,j) - A_{G2_{2.5D}}(i,j)|^2} \tag{4.26}$$

The subject of the sample template with which the query sample template yields minimum β is deemed as the identity of the query sample.

4.4 Experimental results

We perform experiments on 2D and 3D PolyU palmprint database [132]. This database contains region of interest (ROI) extracted from 8,000 samples of 400 different subject palms of 200 volunteers. 2D and 3D palmprints of each subject are registered. Among the volunteers, 136 were male and the other 64 were female. 20 samples from each of these palms were collected in two separate sessions, where 10 samples were captured in each session, respectively. The average time interval between the two sessions is one month. In order to evaluate the performance of the proposed system, we use 200 subject samples collected in the first session as training set in Enrollment stage and corresponding subject samples collected in the second session as the testing set in Identification stage. For the evaluation of the proposed multi-biometric system, we use Equal Error Rate (EER) and Recognition accuracy as the performance metrics. To evaluate the fusion performance, we use Entropy(E), Standard Deviation(SD) and $SIndex$.

4.4.1 Quality improvement in 2.5D palmprint

In the preprocessing stage of our pipeline we enhance the 2D and 3D palmprints by using guided filtering with same palmprint images as guidance. The enhancement in terms of contrast, illumination, principal and wrinkle lines features due to the guided filtering can be observed in Fig. 4.4 (c) and (d). Improvement in sharpness, denoising with texture and depth information preservation as a result of signal level fusion is shown in Fig. 4.4 (e). We also show the mesh plots of 2D, 3D and 2.5D fused palmprint data to illustrate the information integration (Fig. 4.3 (a-c)). Improvement in 2.5D data stems from successful integration of coarser texture information of 2D and finer texture information and depth information of 3D palmprint data. Table 4.2 quantitatively corroborates our claim that 2.5D data has indeed better information content over that of 2D and 3D.

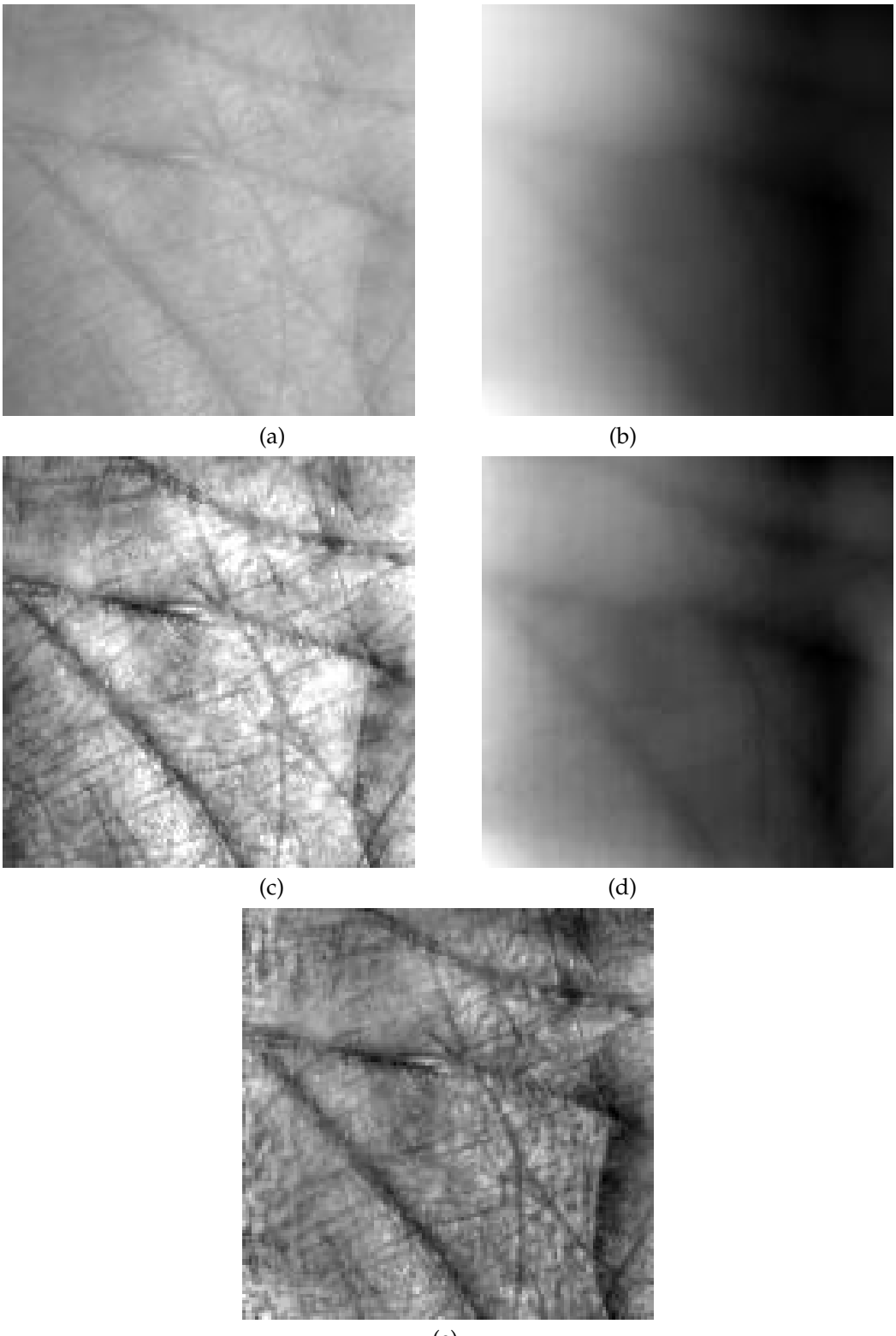


FIGURE 4.4: Illustration: Improvement in quality of palmprints using guided filtering and 2.5D palmprint: (a),(b) Original 2D and 3D palmprints, (c),(d)- Guided filtered 2D and 3D palmprints, (e) Fused 2.5D palmprint.

TABLE 4.2: Average improvement in 2.5D palmprint over 2D and 3D palmprints: In terms of Entropy E, Standard Deviation (SD), $SIndex$.

	E	SD	$SIndex$
2D Palmprint	6.0094	0.0714	3.9992
3D Palmprint	1.8143	3.0693	0.3953
2.5D Palmprint	7.0483	0.1285	3.7652

4.4.2 EER and Recognition accuracy

We use 10 samples ($P = 10$) from each of 200 subjects ($N = 200$) acquired in the first session for the training purpose. To obtain the verification accuracy in terms of EER, each sample (2D and 3D palmprint) is matched with all the samples in the database, resulting in $\binom{P}{2} * N$=9000 genuine and $\binom{N*P}{2}$=1990000 imposter matching scores. Fig. 4.5 shows the Genuine and Imposter normalized frequencies versus the matching score plots. The threshold of the Decision Module was selected at a tradeoff between the false acceptance rate (FAR) and the false rejection rate (FRR). This can be achieved by setting the threshold to the operating point (189.46) at which Genuine and Imposter distributions cross each-other as shown in Fig. 4.5. The performance of the proposed method is compared with some existing 2D and 3D palmprint based recognition systems [26, 31, 32]. As shown in Table 4.3 the EER of the proposed system is clearly below (better) than most of the methods in [26, 31, 32]. It is just marginally above (worse than) the score level fusion of [26]. But, in [26] authors have used pixel level Gabor features which has proved to be computationally complex due to its non-orthogonality and use of number of orientations in the analysis [145].

We also conduct an experiment to derive the recognition accuracy of our system. The recognition accuracy is the percentage of correctly identified Genuine and Imposters by the system. Here, we use 10 samples of all 100 subjects acquired from the second session as probe samples. In [26, 31], the identification accuracy is not reported. Our proposed system gives identification accuracy of 98%. This value is superior than the identification accuracies of 90.69% obtained using only ST feature and 97.38% obtained using only Haralick features. These results clearly justify our choice

of features which in turn points to the merit of constructing a block based graph construction and a 2.5D palmprint data in the first place. Fig. 4.6 shows Receiver Operating Characteristic (ROC) curves of the proposed recognition system and the best performing methods from [31] like- 2D only, 3D only (MW), 2D+3D SVM score level fusion and 2D+3D feature level fusion are plotted. To have fair comparison, we have enlarged the region of interest of ROC curves enclosed in ellipse and given in Fig. 4.7. The figures clearly demonstrate that the proposed 2.5D palmprint based recognition system achieves higher accuracy.

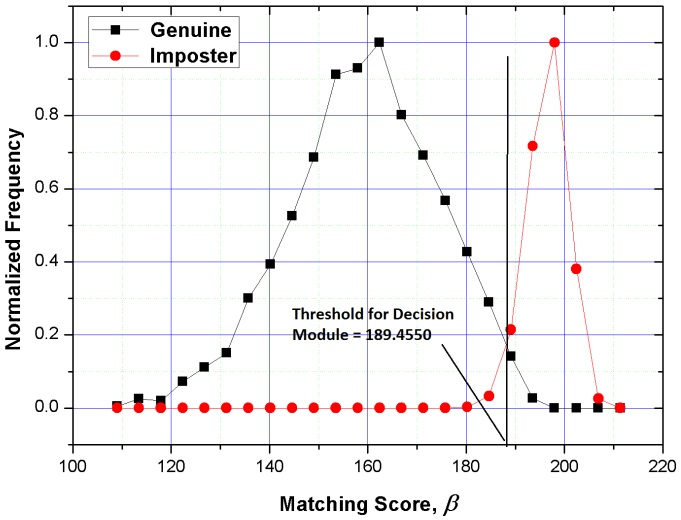
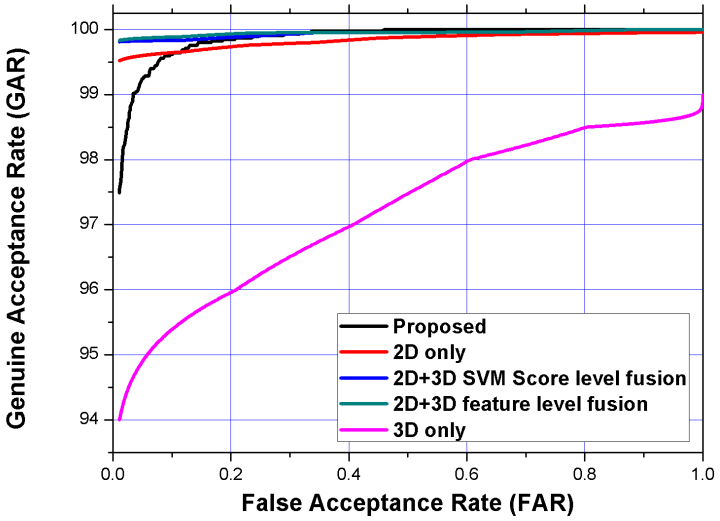


FIGURE 4.5: Selection of optimum Threshold for Decision module

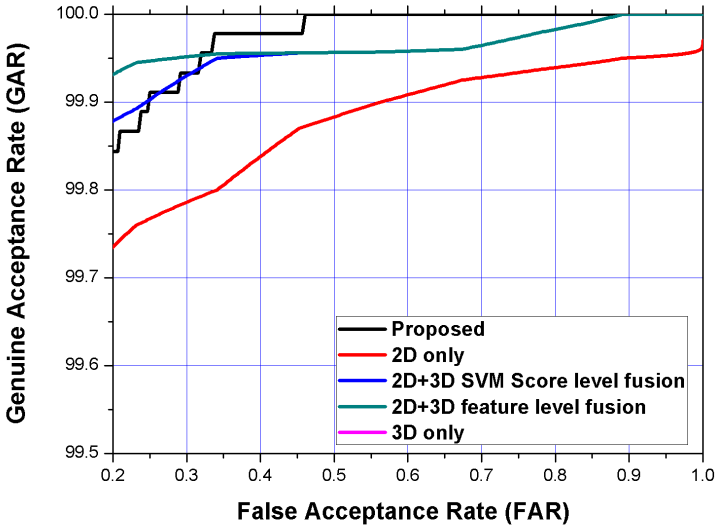
TABLE 4.3: Performance in terms of EER

Recognition System	EER
2D [26]	0.0621
3D [26]	0.9914
Multi-level 2D+3D [26]	**0.0022**
2D+3D SVM Score level fusion [31]	0.0680
2D+3D feature level fusion[31]	0.0590
Joint 2D and 3D matching [32]	0.0250
(2D+3D) method in [146]	0.5600
Proposed method with only ST features	0.0707
Proposed method with only Haralick features	0.0213
Proposed method (with Haralick + ST features)	**0.0179**



centering

FIGURE 4.6: ROC curves of Proposed, 2D palmprint, 3D palmprint, 2D+3D SVM score level fusion, 2D+3D feature level fusion [31].



centering

FIGURE 4.7: Enlargement of region of interest of ROC curves shown in Fig. 4.6.

4.5 Discussions

In this third work we proposed a mult-biometric system based on 2D and 3D palmprint traits. Comprehensive comparisons with several similar approaches on PolyU palmprint database [132] clearly show the superiority of the proposed recognition system. We attribute the improvement to our signal level fusion and graph based

matching scheme. In future, we plan to propose more robust feature extraction from 2.5D data for further improvement in the recognition performance. Another option for further research could be the incorporation of additional traits like IRIS with 2.5D palmprint to make the system more robust and accurate.

Chapter 5

Conclusions and Future Directions

In the first section of this chapter, we conclude our work with highlighting the key contributions made. Certain potential future directions of research are presented in the second part of the chapter.

5.1 Concluding Remarks

The goal of multispectral image or video fusion is to combine complementary information from multiple sources to provide a more informative, complete and accurate representation. It has very widespread applications in diverse fields like medicine, surveillance, military and law enforcement, remote sensing, biometrics, manufacturing, intelligent robots. Over the last decade the advancement in multispectral sensor technology and availability of high processing power at very affordable cost motivated the researchers in the computer vision community to contribute in this area.

In this thesis we have focused on the problem of multispectral image and video fusion with an application in biometry using classical image processing and graph-theoretic solutions. In Chapter 2, we addressed a problem on multispectral image fusion. We present a new focus measure based on steerable local frequency (SLF). The proposed focus measure is shown to perform well in different spectra. Better performance of the proposed focus measure is due to the use of orientation selective local frequency in the source images. We further demonstrated that the proposed focus

measure improves mutispectral multifocus fusion. In the visual spectrum, our fusion scheme outperforms some of the robust and efficient multiresolution transform based methods in addition to some IPD based approaches. In the near-infrared spectrum the proposed fusion method offers a decent performance in comparison with the spatial and transform domain based approaches. In the thermal spectrum, the results show significant improvement over previously reported results. Further, to achieve better fusion performance we proposed a solution based on guided steerable frequency and improved Saliency (GSLF-IS). Superior fusion results are achieved by combining guided steerable local frequency (GSLF) maps with the improved saliency (IS) maps over all spectra.

In the Chapter 3, we worked on multispectral causal video fusion. First, we proposed causal video segmentation method using superseeds and graph matching. The proposed causal video segmentation algorithm surpasses all the existing methods. In the next part we proposed a novel superpixel based causal multisensor video fusion method (CMVF). Here we propose an efficient superpixel level spatio-temporal saliency model as well as superpixel level fusion rules. Comprehensive comparison with several existing approaches on a number of publicly available datasets clearly indicate the advantage of our fusion method.

In Chapter 4, we provide an application of fusion for multi-biometric recognition. We present a solution on multimodal biometric authentication based on fusion of 2D and 3D palmprints. Comparisons with very recent score and feature level fusion based approaches shows the superiority of the proposed system. We attribute the improvement to our signal/low level fusion and novel graph based template generation and matching scheme.

5.2 Future directions

Image and video fusion area has tremendous potential and prospect for the near future. New advancements in multispectral sensor technology and availability of high processing power at very affordable cost are motivating the computer vision community to contribute constantly in this area. In this thesis, we have proposed novel solutions for multispectral image and video fusion.

In future, we plan to extend pixel based fusion to region or object level for further enhancement of multispecral multifocus image fusion. Also importance measures based on intensity/color information need to be exploited to obtain a better focus measure which in turn would also improve the fusion results. In case of multispectral video fusion problem we will examine if superpixel extraction can be made faster which in turn would further reduce the execution time of the proposed causal video fusion algorithm. Another direction of future research will be to analyze the fused video to solve important surveillance tasks like anomalous event detection [8] and person re-identification [9]. We also plan to work on extraction of more robust features from 2.5D palmprint data for improvement in the recognition accuracy of our proposed multi-biometric system. Here, another option for further research could be use of another trait like IRIS with 2.5D palmprint to make the system more robust and accurate.

Appendix A

Basis and interpolation functions of steerable quadrature pair

We have included Tables I and II from [59] which were used for the computation of the oriented analytic image.

TABLE A.1: X-Y seperable basis set and interpolation functions for fourth derivatives of gaussian.

$G_4a = 1.246(0.75 - 3x^2 + x^4)e^{-(x^2+y^2)}$	$K_a(\theta) = \cos^4(\theta)$
$G_4b = 1.246(-1.5x + x^3)(y)e^{-(x^2+y^2)}$	$K_b(\theta) = -4\cos^3(\theta)\sin(\theta)$
$G_4c = 1.246(x^2 - 0.5)(y^2 - 0.5)e^{-(x^2+y^2)}$	$K_c(\theta) = 6\cos^2(\theta)\sin^2(\theta)$
$G_4d = 1.246(-1.5y + y^3)(x)e^{-(x^2+y^2)}$	$K_d(\theta) = -4\cos(\theta)\sin^3(\theta)$
$G_4e = 1.246(0.75 - 3y^2 + y^4)e^{-(x^2+y^2)}$	$K_e(\theta) = \sin^4(\theta)$

TABLE A.2: X-Y seperable basis set and interpolation functions fit for hilbert transform of fourth order derivative of gaussian.

$H_4a = 0.3975(7.189x - 7.501x^3 + x^5)e^{-(x^2+y^2)}$	$K_a(\theta) = \cos^5(\theta)$
$H_4b = 0.3975(1.438 - 4.501x^2 + x^4)(y)e^{-(x^2+y^2)}$	$K_b(\theta) = -5\cos^4(\theta)\sin(\theta)$
$H_4c = 0.3975(x^3 - 2.225x)(y^2 - 0.6638)e^{-(x^2+y^2)}$	$K_c(\theta) = 10\cos^3(\theta)\sin^2(\theta)$
$H_4d = 0.3975(y^3 - 2.225y)(x^2 - 0.6638)e^{-(x^2+y^2)}$	$K_d(\theta) = -10\cos^2(\theta)\sin^3(\theta)$
$H_4e = 0.3975(1.438 - 4.501y^2 - y^4)(x)e^{-(x^2+y^2)}$	$K_e(\theta) = 5\cos(\theta)\sin_{4(\theta)}$
$H_4e = 0.3975(7.189y - 7.501y^3 - y^5)e^{-(x^2+y^2)}$	$\mathrm{K}_e(\theta) = -\sin^5(\theta)$

Bibliography

[1] Z. Wang, D. Ziou, C. Armenakis, D. Li, and Q. Li, "A comparative analysis of image fusion methods", *IEEE Trans. Geosciences and Remote Sensing*, vol. 43, pp. 1391–1402, 2005.

[2] R. S. Blum and Z. Liu, *Multi-Sensor Image Fusion and Its Applications*. CRC Press, 2006, ISBN: 9780849334177.

[3] O. Rockinger, "Image sequence fusion using a shift invariant wavelet transform", in *IEEE International Conference on Image Processing*, Santa Barbara CA, 1997, pp. 288–291.

[4] S. Denman, T. Lamb, C. Fookes, V. Chandran, and S. Sridharan, "Multi-spectral fusion for surveillance systems", *Comput. Electric. Eng.*, vol. 36, no. 4, pp. 643–663, 2010.

[5] R. Raghavendra, B. Dorizzi, A. Rao, and G. H. Kumar, "Designing efficient fusion schemes for multimodal biometric systems using face and palmprint", *Pattern Recognition*, vol. 44, no. 5, pp. 1076–1088, 2011.

[6] T. Stathaki, *Image Fusion: Algorithms and applications*. London, UK: Academic Press, 2008, ISBN: 0123725291, 9780123725295.

[7] A. A. Goshtasby and S. G. Nikolov, "Image fusion: advances in the state of the art", *Information Fusion*, vol. 8, no. 2, pp. 114–118, 2007.

[8] S. Wu, H.-S. Wong, and Z. Yu, "A bayesian model for crowd escape behavior detection", *IEEE Trans. Circuits Syst. Video Techn.*, vol. 24, no. 1, pp. 85–98, 2014.

[9] N. Martinel, S. Abir Das, C. Micheloni, and A. K. Roy-Chowdhury, "Re-identification in the function space of feature warps", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 8, pp. 1656–1669, 2015.

[10] E. P. Bennett, J. L. Mason, and L. McMillan, "Multispectral bilateral video fusion", *IEEE Transactions on Image Processing*, vol. 16, no. 5, pp. 1185–1194, 2007.

[11] N. D. Rasmussen, B. S. Morse, M. A. Goodrich, and D. Eggett, "Fused visible and infrared video for use in wilderness search and rescue", in *Proceedings of IEEE Applications of Computer Vision (WACV), 2009 Workshop on*, Snowbird, UT, 2009, pp. 1–8.

[12] Q. Zhang, L. Wang, Z. Ma, and H. Li, "A novel video fusion framework using surfacelet transform", *Opt. Commun*, vol. 285, no. 13, 14, pp. 3032–3041, 2012.

[13] Q. Zhang, Y. Chen, and L. Wang, "Multisensor video fusion based on spatial–temporal salience detection", *Signal Processing*, vol. 93, pp. 2485–2499, 2013.

[14] Q. Zhang, Y. Wang, M. D. Levine, X. Yuan, and L. Wang, "Multisensor video fusion based on higher order singular value decomposition", *Informaion Fusion*, vol. 24, pp. 54–71, 2015.

[15] M. M. Monwar and M. L. Gavrilova, "Multimodal biometric system using rank-level fusion approach", *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 39, no. 4, pp. 867–878, 2009.

[16] A. Ross and A. K. Jain, "Multimodal biometrics: an overview", in *Proc. of 12th European Signal Processing Conference (EUSIPCO*, 2004.

[17] A. Ross, K. Nandakumar, and A. Jain, *Handbook of Multibiometrics*. New York: Springer-Verlag, 2006, ISBN: 978-0-387-33123-2.

[18] A. Ross, "An introduction to multibiometrics", in *Proc. of the 15th European Signal Processing Conference (EUSIPCO)*, Poznan, Poland, 2007.

[19] A. K. Jain and A. Ross, "Fingerprint mosaicking", in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, Orlando, Florida, USA, 2002, pp. 4064–4067.

[20] J. A. Unar, W. C. Seng, and A. Abbasi, "A review of biometric technology along with trends and prospects", *Pattern Recognition*, vol. 47, no. 8, pp. 2673–2688, 2014.

[21] E. Krotkov, "Focussing", *Computer Vision*, vol. 1, pp. 223–237, 1987.

[22] M. Subbarao, T. Choi, and A. Nikzad, "Focussing techniques", *Optical Engineering*, vol. 32, pp. 2824–2836, 1993.

[23] M. Zukal, J. Mekyska, P. Cika, and Z. Smekal, "Interest points as a focus measure in multispectral imaging", *Radioengineering*, vol. 22, no. 1, pp. 68–81, 2013.

[24] T. D. Dixon, S. G. Nikolov, J. J. Lewis, J. Li, E. F. Canga, J. M. Noyes, T. Troscianko, D. R. Bull, and C. N. Canagarajah, "Task-based scanpath assessment of multi-sensor video fusion in complex scenarios", *Information Fusion*, vol. 11, pp. 51–65, 2010.

[25] L. Xu, J. Du, and Z. Zhang, "Infrared-visible video fusion based on motion compensated wavelet transforms", *IET Image Processing*, vol. 9, no. 4, pp. 318–328, 2015.

[26] D. Zhang, V. Kanhangad, N. Luo, and A. Kumar, "Robust palmprint verification using 2d and 3d features", *Pattern Recognition*, vol. 43, no. 1, pp. 358–368, 2010.

[27] X. Lu, A. K. Jain, and D. Colbry, "Matching 2.5d face scans to 3d models", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 1, pp. 31–43, 2006.

[28] H. Chen and B. Bhanu, "Human ear recognition in 3d", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 4, pp. 718–737, 2007.

[29] P. Yan and K. W. Bowyer, "Biometric recognition using 3d ear shape", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 8, pp. 1297–1308, 2007.

[30] A. K. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems", *Pattern Recognition*, vol. 38, no. 12, pp. 2270–2285, 2005.

[31] D. Zhang, G. Lu, W. Li, L. Zhang, and N. Luo, "Palmprint recognition using 3-d information", *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, vol. 39, no. 5, pp. 505–519, 2009.

[32] W. Li, L. Zhang, D. Zhang, G. Lu, and J. Yan, "Efficient joint 2d and 3d palmprint matching with alignment refinement", in *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, 2010, pp. 795–801.

[33] A. Meraoumia, S. Chitroub, and A. Bouridane, "2d and 3d palmprint information and hidden markov model for improved identification performance", in *11th International Conference on Intelligent Systems Design and Applications, ISDA 2011, Córdoba, Spain, November 22-24, 2011*, 2011, pp. 648–653.

[34] A. Santos, C. Ortiz de Solorzano, J. J. Vaquero, J. M. Peña, N. Malpica, and F. Del Pozo, "Evaluation of autofocus functions in molecular cytogenetic analysis", *Journal of microscopy*, vol. 188, no. 3, pp. 264–272, Dec. 1997.

[35] J. Brenner, B. Dew, J. Horton, J. King, P. Neirath, and S. W., "An automated microscope for cytololologic research", *J. histochem Cytochem*, vol. 24, pp. 100–111, 1971.

[36] T. T. E. Yeo, S. H. Ong, Jayasooriah, and R. Sinniah, "Autofocusing for tissue microscopy", *Image Vision Comput.*, vol. 11, no. 10, pp. 629–639, 1993.

[37] S. Li, B. Yang, and J. Hu, "Performance comparison of different multi-resolution transforms for image fusion", *Information Fusion*, vol. 12, no. 2, pp. 74–84, 2011.

[38] S. K. Nayar and Y. Nakagawa, "Shape from focus", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 8, pp. 824–831, 1994.

[39] F. C. A. Groen, I. T. Young, and G. Ligthart, "A comparison of different focus functions for use in autofocus algorithms", *Cytometry*, vol. 6, pp. 81–91, 1985.

[40] D. Vollath, "Automatic focusing by correlative methods", *Journal of Microscopy*, vol. 147, no. 3, pp. 279–288, 1987.

[41] ——, "The influence of the scene parameters and of the noise on the behavior of automatic focusing algorithms", *Journal of Microscopy*, vol. 151, pp. 133–146, 1988.

[42] L. Firestone, K. Cook, K. Culp, N. Talsania, and J. Kendall Preston, "Comparison of autofocus methods for automated microscopy", *Cytometry*, vol. 12, pp. 195–206, 1991.

[43] M. L. Mendelsohn and B. H. Mayall, "Computer-oriented analysis of human chromosomes—iii. focus", *Computers in Biology and Medicine*, vol. 2, no. 2, pp. 137 –150, 1972.

[44] S. Li, J. T. Kwok, and Y. Wang, "Combination of images with diverse focuses using the spatial frequency", *Information Fusion*, vol. 2, pp. 169–176, 2001.

[45] G. Yang and B. J. Nelson, "Wavelet-based autofocusing and unsupervised segmentation of microscopic images", in *2003 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Las Vegas, Nevada, USA, 2003, pp. 2143–2148.

[46] ——, "Micromanipulation contact transition control by selective focusing and microforce control", in *Proceedings of the 2003 IEEE International Conference on Robotics and Automation, ICRA*, Taipei, Taiwan, 2003, pp. 3200–3206.

[47] J. Tian and L. Chen, "Adaptive multi-focus image fusion using a wavelet-based statistical sharpness measure", *Signal Processing*, vol. 92, no. 9, pp. 2137–2146, 2012.

[48] J. Mekyska, M. Zukal, P. Cika, and Z. Smékal, "Interest points as a focus measure", in *35th International Conference on Telecommunications and Signal Processing, TSP*, Prague, Czech Republic, 2012, pp. 774–778.

[49] J. F. Khan, S. M. A. Bhuiyan, and R. R. Adhami, "Feature point extraction from the local frequency map of an image", *J. Electrical and Computer Engineering*, vol. 2012, 182309:1–182309:15, 2012.

[50] C. Wu and Q. Wang, "A novel approach for interest point detection based on phase congruency", in *IEEE TENCON Conference*, Melbourne, Qld, 2005, pp. 1–6.

[51] S. Das and M. K. Kundu, "A neuro-fuzzy approach for medical image fusion", *IEEE Trans. Biomed. Engineering*, vol. 60, no. 12, pp. 3347–3353, 2013.

[52] R. Benes, P. Dvorak, M. Faúndez-Zanuy, V. Espinosa-Duro, and J. Mekyska, "Multi-focus thermal image fusion", *Pattern Recognition Letters*, vol. 34, no. 5, pp. 536–544, 2013.

[53] R. Minhas, A. A. Mohammed, and Q. M. J. Wu, "An efficient algorithm for focus measure computation in constant time", *IEEE Trans. Circuits Syst. Video Techn.*, vol. 22, no. 1, pp. 152–156, 2012.

[54] A. Oppenheim and J. S. Lim, "The importance of phase in signals", *Proceedings of the IEEE*, vol. 69, pp. 529–541, 1981.

[55] X. Y. Liu, W. H. Wang, and Y. Sun, "Dynamic evaluation of autofocusing for automated microscopic analysis of blood smear and pap smear", *J. Microsc.*, vol. 227, pp. 15–23, 2007.

[56] H. Zhao, Z. Shang, Y. Y. Tang, and B. Fang, "Multi-focus image fusion based on the neighbor distance", *Pattern Recognition*, vol. 46, no. 3, pp. 1002–1011, 2013.

[57] M. Faundez-Zanuya, J. Mekyska, and V. Espinosa-Duró, "On the focusing of thermal images", *Pattern Recognition Letters*, vol. 32, no. 11, pp. 1548–1557, 2011.

[58] F. Crete, T. Dolmiere, P. Ladret, and M. Nicolas, "Perception and estimation with a new no-reference perceptual blur metric", in *Proc. SPIE 6492, Human Vision and Electronic Imaging XII*, San Jose, CA, USA, 2007, 64920I.1—64920I.11.

[59] W. T. Freeman and E. H.Adelson, "The design and use of steerable filters", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 9, pp. 891–906, 1991.

[60] P. Danielsson and O. Seger, "Rotation invariance in gradient and higher order derivative detectors", *Computer Vision Graphics Image Processing*, vol. 49, no. 2, pp. 198–221, 1990.

[61] W. T. Freeman and E. H.Adelson, "Steerable filters", *Topical Mtg. Image Understanding Machine Vision Opt. Soc. Amer.,Tech. Digest Series*, vol. 14, 1989.

[62] U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is bottom-up attention useful for object recognition?", in *Proceedings of IEEE Comp. Soc. Conf. Computer Vision and Pattern recognition*, Washington, DC, USA, 2004, pp. 37–44.

[63] H. Li and K. N. Ngan, "A co-saliency model of image pairs", *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3365–3375, 2011.

[64] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency", in *Proceedings of Neural Information Processing Systems*, Vancouver, British Columbia, Canada, 2006, pp. 545–552.

[65] J. Malik and P. Perona, "Preattentive texture discrimination with early vision mechanisms", *Journal of the Optical society of America A*, vol. 7, no. 5, pp. 923–932, 1990.

[66] 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.

[67] X. Hou and L. Zhang, "Saliency detection: a spectral residual approach", in *Proceedings of IEEE Comp. Soc. Conf. Computer Vision and Pattern recognition*, Minneapolis, Minnesota, USA, 2007.

[68] K. He, J. Sun, and X. Tang, "Guided image filtering", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1397–1409, 2013.

[69] ——, "Guided image filtering", in *Proceedings of 11th European Conference on Computer Vision (ECCV 2010)*, Heraklion, Crete, Greece, 2010, pp. 1–14.

[70] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images", in *Proceedings of IEEE International Computer Vision Conference (ICCV 1998)*, Bombay, India, 1998, pp. 839–846.

[71] G. Petschnigg, R. Szeliski, M. Agrawala, M. F. Cohen, H. Hoppe, and K. Toyama, "Digital photography with flash and no- flash image pairs", *ACM Trans. Graph.*, vol. 23, no. 3, pp. 664–672, 2004.

[72] S. Chaudhuri, S. Chatterjee, N. Katz, M. Nelson, and M. Goldbaum, "Detection of blood vessels in retinal images using two-dimensional matched filters", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, pp. 263–269, 1989.

[73] A. Saha, G. Bhatnagar, and Q. M. J. Wu, "Mutual spectral residual approach for multifocus image fusion", *Digital Signal Processing*, vol. 23, no. 4, pp. 1121–1135, 2013.

[74] Q. Guihong, Z. Dali, and P. Yan, "Information measure for performance of image fusion", *Electr. Letters*, vol. 38(7), pp. 313–315, 2002.

[75] C. S. Xydeas and V. Petrović, "Objective image fusion performance measure", *Electr. Letters*, vol. 36, pp. 308–309, 2000.

[76] Z. Wang and A. C. Bovik, "A universal image quality index", *IEEE Signal Processing Letters*, vol. 9(3), pp. 81–84, 2002.

[77] W. Xiuqing, Z. Rong, and X. Yunxiang, "A method of wavelet-based edge detection with data fusion for multiple images", in *Proceedings of the 3rd world Congress on Intelligent Control and Automation*, China, 2000, pp. 2691–2694.

[78] I. Haritaoglu, D. Harwood, and L. S. Davis, "Real-time surveillance of people and their activities", vol. 22, pp. 781–796, 2000.

[79] Z. Zhang and R. S. Blum, "A categorization of multiscale-decomposition-based image fusion schemes with a performance study for a digital camera application", in *Proceedings of the IEEE*, vol. 87, 1999, pp. 1315–1326.

[80] J. Xu, S. Denman, S. Sridharan, and C. Fookes, "An efficient and robust system for multiperson event detection in real-world indoor surveillance scenes", *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 25, no. 6, pp. 1063–1076, 2015.

[81] P. Foggia, A. Saggese, and M. Vento, "Real-time fire detection for video-surveillance applications using a combination of experts based on color, shape, and motion", *IEEE Trans. Circuits Syst. Video Techn.*, vol. 25, no. 9, pp. 1545–1556, 2015.

[82] S. Wei, Y. Zhao, C. Zhu, C. Xu, and Z. Zhu, "Frame fusion for video copy detection", *IEEE Trans. Circuits Syst. Video Techn.*, vol. 21, no. 1, pp. 15–28, 2011.

[83] X. Wei, Y. Jiang, and C. Ngo, "Concept-driven multi-modality fusion for video search", *IEEE Trans. Circuits Syst. Video Techn.*, vol. 21, no. 1, pp. 62–73, 2011.

[84] A. Briassouli and N. Ahuja, "Fusion of frequency and spatial domain information for motion analysis", in *Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)*, vol. 2, Cambridge, UK, 2004, pp. 175–178.

[85] A. Torabi, G. Masse, and G.-A. Bilodeau, "An iterative integrated framework for thermal-visible image registration, sensor fusion, and people tracking for video survellaince applications", *Computer vision and Image Understanding*, vol. 116, pp. 210–221, 2012.

[86] S. S. Pillai and M. Swamy, "Camouflaged target detection using real-time video fusion algorithm based on multi-scale transforms", in *Electrical and Computer Engineering (CCECE), 2014 IEEE 27th Canadian Conference on*, Toronto, ON, May 2014, pp. 1–5.

[87] C. Couprie, C. Farabet, Y. LeCun, and L. Najman, "Causal graph-based video segmentation", in *IEEE International Conference on Image Processing, ICIP*, Melbourne, Australia, 2013, pp. 4249–4253.

[88] F. Galasso, R. Cipolla, and B. Schiele, "Video segmentation with superpixels", in *Computer Vision - ACCV 2012 - 11th Asian Conference on Computer Vision, Revised Selected Papers, Part I*, Daejeon, Korea, 2012, pp. 760–774.

[89] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, 2012.

[90] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, 2002.

[91] Y. J. Lee, J. Kim, and K. Grauman, "Key-segments for video object segmentation", in *IEEE International Conference on Computer Vision, ICCV 2011*, Barcelona, Spain, 2011, pp. 1995–2002.

[92] M. P. Kumar, P. H. S. Torr, and A. Zisserman, "Learning layered motion segmentations of video", *International Journal of Computer Vision*, vol. 76, no. 3, pp. 301–319, 2008.

[93] K. J. F. de Souza, A. de Albuquerque Araújo, Z. K. G. do Patrocínio Jr., and S. J. F. Guimarães, "Graph-based hierarchical video segmentation based on a simple dissimilarity measure", *Pattern Recognition Letters*, vol. 47, pp. 85–92, 2014.

[94] M. Grundmann, V. Kwatra, M. Han, and I. A. Essa, "Efficient hierarchical graph-based video segmentation", in *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, San Francisco, CA, USA, 2010, pp. 2141–2148.

[95] F. Galasso, M. Iwasaki, K. Nobori, and R. Cipolla, "Spatio-temporal clustering of probabilistic region trajectories", in *IEEE International Conference on Computer Vision, ICCV 2011*, Barcelona, Spain, 2011, pp. 1738–1745.

[96] O. Miksik, D. Munoz, J. A. Bagnell, and M. Hebert, "Efficient temporal consistency for streaming video scene analysis", in *2013 IEEE International Conference on Robotics and Automation*, Karlsruhe, Germany, 2013, pp. 133–139.

[97] S. Paris, "Edge-preserving smoothing and mean-shift segmentation of video streams", in *Computer Vision - ECCV 2008, 10th European Conference on Computer Vision, Proceedings, Part II*, Marseille, France, 2008, pp. 460–473.

[98] Y. Zhou, X. Bai, W. Liu, and L. J. Latecki, "Fusion with diffusion for robust visual tracking", in *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting.*, Lake Tahoe, Nevada, USA, 2012, pp. 2987–2995.

[99] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise", in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, Portland, Oregon, USA, 1996, pp. 226–231.

[100] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, 2002.

[101] D. Koutra, A. Parikh, A. Ramdas, and J. Xiang, "Algorithms for graph similarity and subgraph matching.", School of Computer Science, Carneige Mellon university, Pittsburgh, PA, Tech. Rep., 2011. [Online]. Available: //www.cs.cmu.edu/~jingx/docs/DBreport.pdf..

[102] F. Meyer, "Topographic distance and watershed lines", *Signal Process.*, vol. 38, no. 1, pp. 113–125, 1994.

[103] J. B. Roerdink and A. Meijster, "The watershed transform: definitions, algorithms and parallelization strategies", *Fundam. Inf.*, vol. 41, no. 1,2, pp. 187–228, Apr. 2000.

[104] G. Csurka, D. Larlus, and F. Perronnin, "What is a good evaluation measure for semantic segmentation?", in *British Machine Vision Conference, BMVC*, Bristol, UK, 2013.

[105] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images", in *Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision Proceedings, Part V*, Florence, Italy, 2012, pp. 746–760.

[106] Y. Wo, X. Chen, and G. Han, "A saliency detection model using aggregation degree of color and texture", *Signal Processing:Image Communication*, vol. 30, pp. 121–136, 2015.

[107] H. Jing, X. He, Q. Han, A. A. A. El-Latif, and X. Niu, "Saliency detection based on integrated features", *Neurocomputing*, vol. 129, pp. 114–121, 2014.

[108] W. Kim and J.-J. Han, "Video saliency detection using contrast of spatiotemporal directional coherence", vol. 21, no. 10, pp. 1250–1254, 2014.

[109] Z. Liu, X. Zhang, S. Luo, and O. L. Meur, "Superpixel-based spatiotemporal saliency detection", *IEEE Trans. Circuits Syst. Video Techn.*, vol. 24, no. 9, pp. 1522–1540, 2014.

[110] Y. Wang, J. See, R. C.-W. Phan, and Y.-H. Oh, "Efficient spatio-temporal local binary patterns for spontaneous facial micro-expression recognition", *PLoS ONE*, vol. 10, no. 5, pp. 1–20, 2015.

[111] J. W. Davis and V. Sharma, "Background-subtraction using contour-based fusion of thermal and visible imagery", *Computer Vision and Image Understanding*, vol. 106, no. 2-3, pp. 162–182, 2007.

[112] C. O. Conaire. (2015). Aic thermal/visible night-time dataset, [Online]. Available: http://www.eeng.dcu.ie/~oconaire/dataset/.

[113] J. J. Lewis, S. G. Nikolov, A. Loza, E. F. Canga, N. Cvejic, J. Li, A. Cardinali, C. N. Canagarajah, D. R. Bull, T. Riley, D. Hickman, and M. I. Smith, "The eden project multi-sensor data set", University of Bristol and Waterfall Solutions Ltd, UK, Tech. Rep. TR-UoB-WS-Eden-Project-Data-Set, 2006.

[114] Y. Benezeth, D. Sidibé, and J.-B. Thomas, "Background subtraction with multispectral video sequences", in *IEEE International Conference on Robotics and Automation workshop on Non-classical Cameras, Camera Networks and Omnidirectional Vision (OMNIVIS)*, Hong Kong SAR China, Jun. 2014, 6 p.

[115] J. Shen, Y. Du, W. Wang, and X. Li, "Lazy random walks for superpixel segmentation", *IEEE Trans. Image Processing*, vol. 23, no. 4, pp. 1451–1462, 2014.

[116] G. Piella and H. Heijmans, "New quality measures for image fusion", in *Proceedings of the 7th International Conference on Information Fusion*, Stockholm, Sweden, 2004, pp. 542–546.

[117] V. Petrovic, T. Cootes, and R. Pavlovic, "Dynamic image fusion performance evaluation", in *Proceedings of the IEEE International Conference on Information Fusion*, Quebec, Que., 2007, pp. 1–7.

[118] E. H. Adelson, C. H. Anderson, J. R. Bergen, P. J. Burt, and J. Qgden, "Pyramid methods in signal processing", *RCA Eng*, vol. 29, pp. 33–41, 1984.

[119] J. Liang, Y. He, D. Liu, and X. Zeng, "Image fusion using higher order singular value decomposition", *IEEE Transactions on Image Processing*, vol. 21, no. 5, pp. 2898–2909, 2012.

[120] X. B. Qu, J. W. Yan, H. Z. Xiao, and Z. Q. Zhu, "Image fusion algorithm based on spatial frequency-motivated pulse coupled neural networks in nonsubsampled contourlet transform domain", *Acta Automatica Sinica*, vol. 34, no. 12, pp. 1508–1514, 2008.

[121] W. Li, D. Zhang, and ZhuoqunXu, "Palmprint identification by fourier transform", *IJPRAI*, vol. 16, no. 4, pp. 417–432, 2002.

[122] X. Wu, D. Zhang, and K. Wang, "Fisherpalms based palmprint recognition", *Pattern Recognition Letters*, vol. 24, no. 15, pp. 2829–2838, 2003.

[123] D. Zhang, A. W. Kong, J. You, and M. Wong, "Online palmprint identification", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1041–1050, 2003.

[124] Z. Sun, T. Tan, Y. Wang, and S. Z. Li, "Ordinal palmprint represention for personal identification", in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA*, 2005, pp. 279–284.

[125] T. Connie, A. T. B. Jin, M. G. K. Ong, and D. N. C. Ling, "An automated palmprint recognition system", *Image Vision Comput.*, vol. 23, no. 5, pp. 501–515, 2005.

[126] L. Shang, D. Huang, J. Du, and Chun-HouZheng, "Palmprint recognition using fast ica algorithm and radial basis probabilistic neural network", *Neurocomputing*, vol. 69, no. 13-15, pp. 1782–1786, 2006.

[127] P. H. Hennings-Yeomans, B. V. K. V. Kumar, and M. Savvides, "Palmprint classification using multiple advanced correlation filters and palm-specific segmentation", *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 3-2, pp. 613–622, 2007.

[128] D. Huang, W. Jia, and D. Zhang, "Palmprint verification based on principal lines", *Pattern Recognition*, vol. 41, no. 4, pp. 1316–1328, 2008.

[129] W. Jia, D. Huang, and D. Zhang, "Palmprint verification based on robust line orientation code", *Pattern Recognition*, vol. 41, no. 5, pp. 1504–1513, 2008.

[130] M. Y. Feng Yue Bin Li and J. Wang, "Hashing based fast palmprint identification for large-scale databases", *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 5, pp. 769–778, 2013.

[131] Y.-K. L. Wei Jia Rong-Xiang Hu, Y. Zhao, and J. Gui, "Histogram of oriented lines for palmprint recognition", *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 44, no. 3, pp. 385–395, 2014.

[132] B. R. centre and T. H. K. P. University. (2015). Polyu palmprint database, [Online]. Available: http://www4.comp.polyu.edu.hk/~biometrics/.

[133] I. A. Kakadiaris, G. Passalis, G. Toderici, M. N. Murtuza, Y. Lu, N. Karampatziakis, and T. Theoharis, "Three-dimensional face recognition in the presence of facial expressions: an annotated deformable model approach", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 4, pp. 640–649, 2007.

[134] C. Samir, A. Srivastava, and M. Daoudi, "Three-dimensional face recognition using shapes of facial curves", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 11, pp. 1858–1863, 2006.

[135] B. Gökberk, H. Dutagaci, A. Ulas, L. Akarun, and B. Sankur, "Representation plurality and fusion for 3-d face recognition", *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 38, no. 1, pp. 155–173, 2008.

[136] D. Z. Adams Kong and M. Kamel, "A survey of palmprint recognition", *Pattern Recognition*, vol. 42,

[137] J. Ni, J. Luo, and W. Liu, "3d palmprint recognition using dempster-shafer fusion theory", *J. Sensors*, vol. 2015, 252086:1–252086:7, 2015.

[138] W. Li, D. Zhang, L. Zhang, G. Lu, and J. Yan, "3-d palmprint recognition with joint line and orientation features", *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, vol. 41, no. 2, pp. 274–279, 2011.

[139] K. He, J. Sun, and X. Tang, "Guided image filtering", in *Proceedings of European Conference on Computer Vision*, 2010.

[140] ——, "Guided image filtering", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1397–1409, 2013.

[141] K. S. S. Robert M. Haralick and I. Dinstein, "Textural features for image classification", *IEEE Transactions Systems, Man, and Cybernetics*, vol. 3, no. 6, pp. 610–621, 1973.

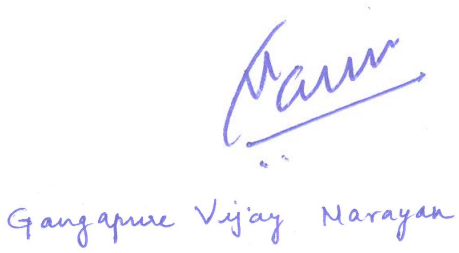
[142] N. Zayed and H. A. Elnemr, "Statistical analysis of haralick texture features to discriminate lung abnormalities", *Biomedical Imaging*, vol. 2015, 267807:1–267807:7, 2015.

[143] P. J. Besl and R. C. Jain, "Segmentation through variable-order surface fitting", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 10, no. 2, pp. 167–192, 1988.

[144] M. P. do Carmo, *Differential geometry of curves and surfaces*. Prentice Hall, 1976, ISBN: 978-0-13-212589-5.

[145] A. T. Gholamreza Amayeh and G. Bebis, "Accurate and efficient computation of gabor features in real-time applications", *Advances in Visual Computing*, vol. 5875, pp. 243–252, 2009.

[146] A. K. Vivek Kanhangad and D. Zhang, "A unified framework for contactless hand verification", *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 3, pp. 1014–1027, 2011.


Gangapure Vijay Narayan