# USE OF COHERENT POINT DRIFT IN COMPUTER VISION APPLICATIONS

**By**

Sara Saravi

A doctoral thesis
Submitted in partial fulfilment of the requirements
for the award of

Doctor of Philosophy

Department of Computer Science

Loughborough University

June 2013



Supervisor: Dr. Eran Edirisinghe

# Abstract

This thesis presents the novel use of Coherent Point Drift in improving the robustness of a number of computer vision applications. CPD approach includes two methods for registering two images - rigid and non-rigid point set approaches – which are based on the transformation model used. The key characteristic of a rigid transformation is that the distance between points is preserved, which means it can be used in the presence of translation, rotation, and scaling. Non-rigid transformations - or affine transforms - provide the opportunity of registering under non-uniform scaling and skew. The idea is to move one point set coherently to align with the second point set. The CPD method finds both the non-rigid transformation and the correspondence distance between two point sets at the same time without having to use a-priori declaration of the transformation model used.

The first part of this thesis is focused on speaker identification in video conferencing. A real-time, audio-coupled video based approach is presented, which focuses more on the video analysis side, rather than the audio analysis that is known to be prone to errors. CPD is effectively utilised for lip movement detection and a temporal face detection approach is used to minimise false positives if face detection algorithm fails to perform.

The second part of the thesis is focused on multi-exposure and multi-focus image fusion with compensation for camera shake. Scale Invariant Feature Transforms (SIFT) are first used to detect keypoints in images being fused. Subsequently this point set is reduced to remove outliers, using RANSAC (RANdom Sample

Consensus) and finally the point sets are registered using CPD with non-rigid transformations. The registered images are then fused with a Countourlet based image fusion algorithm that makes use of a novel alpha blending and filtering technique to minimise artefacts. The thesis evaluates the performance of the algorithm in comparison to a number of state-of-the-art approaches, including the key commercial products available in the market at present, showing significantly improved subjective quality in the fused images.

The final part of the thesis presents a novel approach to Vehicle Make & Model Recognition in CCTV video footage. CPD is used to effectively remove skew of vehicles detected as CCTV cameras are not specifically configured for the VMMR task and may capture vehicles at different approaching angles. A LESH (Local Energy Shape Histogram) feature based approach is used for vehicle make and model recognition with the novelty that temporal processing is used to improve reliability. A number of further algorithms are used to maximise the reliability of the final outcome. Experimental results are provided to prove that the proposed system demonstrates an accuracy in excess of 95% when tested on real CCTV footage with no prior camera calibration.

***Sara Saravi***

***18th June 2013***

To my beloved Parents

***Mahin Dianat*** *and* ***Mohammad Saravi***

# Acknowledgment

This thesis arose in part out of three years of research. During that time, I have worked with a great number of people whose contribution in assorted ways to the research and the making of the thesis deserved special mention. It is a pleasure to convey my gratitude to them all in my humble acknowledgment.

The chain of my gratitude begins with name of God Almighty, the Most Gracious, and the Most Merciful whose blessings have always been with me.

I shall like to express my deep and sincere gratefulness to my supervisor, Prof. Eran Edirisinghe for his supervision, irreplaceable assistance, advice, abundant help and guidance from the very early stage of this research as well as giving me extraordinary experiences throughout the work. His wide knowledge and logical way of thinking have been of great value for me. His detailed and constructive comments, technical and editorial advice was essential to the completion of this thesis. Above all and the most needed, he provided me unwavering encouragement and support in various ways. I am indebted to him more than he knows.

I shall also like to convey thanks to the Computer Science Department, Loughborough University, UK, for providing funding to complete my PhD.

Thanks to my colleagues at Digital Imaging Research Group (DIRG), who shared their knowledge, provided useful feedback on my work, provided a nice and friendly environment to work and made my three years enjoyable and memorable. Special thanks to my friends Dr. Iffat Zafar, Dr.Usman zakir, Dr. Muhammad Athar Ali, Dr. Nesreen Otoum, Mr. M. Akramshah Ismail, Mr. Andrew Leonce, Mr. Niraj Doshi , Ms. Anoud Bani-Hani, Ms. Giounona Tzanidou and Mr. Fraz Fraz for their help, care, support, motivation, and time throughout my PhD. I would also like to thank to my second mother Ms. Christine Bagley for her invaluable support and encouragement. I am also thankful to the technical and clerical staff members of the Department of Computer Science, Loughborough University.

I thank my Iranian friends beginning with my sister Ms. Nazi Zaeri, Mr. Hamid Reza Azizi, Mr. Arash Beizaee, Mr. Amir Badiee and many more for their mental support and encouragements. I am deeply grateful to my flatmate Ms. Ilia Roustemoglu for supporting me with her kindness during stressful times.

It would not be possible to achieve all the success without my family. My parents deserve special mention for their unconditional, inseparable support, prayers and endless love. My Father, Mohammad Saravi and my Mother Mahin Dianat are the ones who put the fundament in my learning character, showing me the joy of intellectual pursuit ever since I was a child. They are the ones who sincerely raised me with their caring and gentle love and prayed day and night for my success. My

special gratitude to my brother: Mehrdad Saravi, my sister-in-law Nazi Saravi and my niece Viyana Saravi. Thanks for being supportive, loving and caring.

***Sara Saravi***

***11th February 2013***

یا هو

سپاس و ستایش خدای را که با نیروی خامه وقلم، انسان را با علم و دانش آشنا ساخت و قدرت شناخت و تفکر و اندیشه را بر او ارزانی داشت.

شکر نعمت او که ره رستگاری و سعادت را بر بستر تحصیل علم برقرار نمود.

درود بر آفریدگاری که پیامبرش، طلب علم را از اوان کودکی تا دم آخر بر همگان واجب نموده است.

حمد بر کردکاری که با انوار جهان تاب و پر فروغش، افق های روشنی در مقابل دیدگان بشر گسترده و نمایان ساخت.

قدر مسلم در وصال به این اهداف متعالی، الطاف الهی آشکار و به صورت های مختلف متجلی میگردد. از آن جمله، حمایت ها وپیگیری های پدر و مادر عزیزم و تشویقات برادر مهربانم که از کرامات غیبی است و دست مایه تحقق این توفیق به حساب می آید، سپاس بیکران و شکرانه بی پایان را طلب می کند.

در مقام شکرگزاری رساله دکترای خود را به پدر و مادر عزیزم، محمد و مهین و خانواده پر مهرم تقدیم میدارم.

سارا ساروی
دوشنبه - ۲۳ بهمن ۱۳۹۱

# Table of Contents

# List of Figures and Tables

## Figures

# Tables

# Abbreviations and Notations

## Abbreviations

| | |
|---|---|
| **ADABOOST** | Adaptive Boosting |
| **CPD** | Coherent Point Drift |
| **DCT** | Discrete Cosine Transform |
| **DFB** | Directional Filter Bank |
| **DoG** | Difference of Gaussian |
| **DWT** | Discrete Wavelet Transform |
| **GMM** | Gaussian Mixture Model |
| **HDR** | High Dynamic Range |
| **HMM** | Hidden Markov Model |
| **LDR** | Low Dynamic Range |
| **LESH** | Local Energy Shape Histogram |
| **LP** | Laplacian Pyramid |
| **MCT** | Motion Coherence Theory |
| **PDF** | Probability Density Function |
| **RANSAC** | RANdom SAmple Consensus |
| **ROI** | Region of Interest |
| **SDR** | Standard Dynamic Range |
| **SIFT** | Scale Invariant Feature Transform |
| **SVM** | Support Vector Machine |
| **VMMR** | Vehicle Make and Model Recognition |
| **WBCT** | Wavelet Based Contourlet Transform |

## Notations

| | |
|---|---|
| $I$ | Image |
| $E$ | Energy |
| $x, y$ | Position of Each Coefficient in Image |
| $X_{N\times D}$ | Reference Point Set, represented by N×D Matrix where D is the dimension of the points. |
| $Y_{M\times D}$ | Template Point Set (GMM centroids), represented by M×D Matrix where D is the dimension of the points. |
| $v$ | Velocity |
| $\lambda$ | Weighting Constant |
| $\phi(v)$ | Regularization Function |
| $\tilde{v}$ | Fourier Transform |
| $v(z)$ | Radial Basis Function |
| $\tilde{G}$ | Symmetric Low-Pass Filter |
| $\sigma$ | Capture Range for Each Gaussian Mixture Component |
| $G$ | Gaussian Kernel |
| $G_{M\times M}$ | Gaussian Affinity Matrix |
| $D$ | Derivative Operator |
| $W_{M\times D}$ | Gaussian Kernel Weights Matrix |
| $Q$ | Upper Bound |
| $P$ | Matrix of Posterior Probabilities |
| $P^{old}$ | Posterior Probabilities Matrix Calculated Using Previous Parameter Values |
| $a$ | Support for Uniform Probability Density Function |
| $\alpha$ | Annealing Rate |
| $\beta$ | Strength of Interactive Between Points |

| | |
|---|---|
| $\lambda$ | Trade-off Between Data Fitting and Smoothness Regularization |
| $G(x, y, \sigma)$ | Gaussian Function |
| $L(x, y, \sigma)$ | Scale Space of Image |
| $k$ | Constant |
| $s, t$ | Scale and Orientation |
| $\psi$ | Gabor Wavelet Kernel Bank |
| $R$ | Response at Image Position x,y |
| $A_n$ | Amplitude |
| $\phi_n$ | Phase of Response |
| $W$ | Weight of Frequency Range |
| $T$ | Estimated Noise Influence |
| $L$ | Orientation Label Map |
| $b$ | Current Bin |
| $r$ | Region |
| $w_r$ | Gaussian Weighting Function |
| $\delta_{Lb}$ | Kronecker Delta |
| $d(p, n)$ | Euclidean Distance Between p,n |
| $E_H$ | High Frequency Sub-band |
| $H = (l, m, n)$ | l: Level of Wavelet Decomposition, m: LH, HL and HH sub-bands of Wavelet Decomposition and n: Directional Contourlet sub-bands |
| $i \times j$ | Rectangular Sub-region |
| $V_L^{(X)}(x, y)$ | Variance of Each Sub-Region |
| $A_L^{(F)}(i, j)$ | Average of Low-pass Wavelet Sub-bands |
| $SB_{out}$ | Sub-band of Fused Block |

# CHAPTER 1

# An Overview

## 1.1 Introduction

In Computer vision the aim is to recognize the world under the cover of a picture. A computer vision system extracts important information from images obtained from a camera and in simple words attempts to simulate the high-dimensional human vision system where the processor is the brain. Therefore Computer vision is introduced as the initiative of automating and assimilating a broad series of developments and demonstrations for vision sensitivity.

One of the applications in computer vision is a system for speaker identification where a video is treated as a sequence of images. Another interesting application is fusing images to create High Dynamic Range (HDR) images. Vehicle Make and Model Recognition is another example of a computer vision application which has a very important usage in access control security systems. A significant amount of research has already being carried out in the above three application areas of computer vision. However these techniques have been limited mostly to algorithms which address the corresponding research issues, partially. This allows an opportunity to contribute to the current open research issues in these application

domains, using the latest advances in computer vision and related fundamental technological areas.

One common issue amongst the application areas mentioned above is challenge faced by standard computer vision algorithms when having to operate accurately in the presence of either camera movement or in the presence of moving objects within scenes captured by stationary cameras. Fundamentally a solution to the above challenge requires the ability to identify and quantify changes in shape of objects analysed so that the information thus obtained can be used to compensate for movement. A detailed review of literature carried out within the context of the research presented in this thesis revealed that Coherent Point Drift (CPD), a novel mathematical theory proposed in the recent past, can be used to study the movement of points within objects of changing shape, when the associated transformation are either rigid or non-rigid.

## 1.2 Research Motivation

The motivations for research covering the context of the investigations presented in this thesis came from a comprehensive literature review in computer vision and related technical and mathematical areas and the subsequent investigation of the shortcomings of the proposed algorithms. These deficiencies of the state-of-the-art technology provided the basic research motivation to seek solutions that utilise different approaches to develop complete, end-to-end solutions to the identified practical problems in the abovementioned computer vision application areas.

The motivation to conduct research to close existing gaps in the three identified practical application domains initiated a further comprehensive literature review to seek the state-of-art in existing solutions. These investigations revealed the following:

1. Although solutions have been proposed to Speaker Identification in videoconferencing, the most effective solution are either based on speech identification or visual information processing to identify high-level facial movements. These approaches lack accuracy and can face significant challenges when faced with addressing likely, practical challenges in a typical videoconferencing event. Initial investigations revealed the ability of Coherent Point Drift to solve this problem and hence provide a more robust solution.
2. Even though in literature a significant amount of research has been conducted and published in the area of multi-exposure and multi-focus image fusion for creating HDR images, detailed investigations carried out within the remits of the research presented in this thesis revealed that the existing algorithms fail in the presence of camera shake. This is a critical shortcoming that makes such algorithms not suitable for implementation in hand held devices. Further it was revealed that even when camera shake was not an issue, displaying the fused images on large screens revealed various artefacts such as, ‘image ghosting’, ‘blockiness’, etc. Our initial investigations showed that Coherent Point Drift can be used as a fundamental approach to registering corresponding point sets required to compensate for camera shake. Further the use of alpha blending performed

appropriately within the multi-resolution Contourlet transform domain was found to be a way that could avert the formation of image artefacts. Therefore it was decided that the above will be investigated within the context of the research presented in this thesis.

3. As security is an increasingly important aspect in day-to-day life, the importance of enhancing the fool proof nature of present state-of-the-art video analytics systems is felt significantly. To this extent, enhancing current Automatic Number Plate Recognition (ANPR) systems with Vehicle Make & Model Recognition (VMMR) systems, has been attempted in the recent past. However all existing approaches have been designed for and have been applied under controlled conditions, assuming that the cameras are specifically set up and configured for the purpose and are rigidly fixed and favourably positioned. Often however there is a need to be able to identify the vehicle make and model based on general purpose (not purposely set and configured) CCTV video footage, which are medium to poor quality and may pick up oncoming vehicles at an angle that makes vehicles appear skewed in shape. These create additional challenges to the VMMR algorithms. As the initial investigations revealed that existing algorithms cannot cater for these challenges. The focus of the latter part of this thesis has been reserved to designing, implementing and testing a VMMR approach that is robust to above and will be able to operate on general purpose CCTV footage.

The above research motivations led to the following aim and objectives.

## 1.3 Aim and Objectives

The **aim** of the research presented in this thesis is to design and develop efficient algorithms for speaker identification, multi-exposure/focus image fusion and vehicle make and model recognition that are robust to a number of practical challenges that often limits present state-of-the-art algorithms being used in practice. In particular facing up to the challenges posed by camera shake/movement, skew and non-linear deformations of objects will be considered.

The specific research **objectives** can be listed below as:

- To carry out literature reviews in areas of research focus of this thesis and to analyse the challenges that the algorithms will face when implemented and used in real systems, supporting practical application tasks.
- To investigate the state-of-art in speaker identification applied to videoconferencing. In particular to investigate the robustness of algorithms in the presence of background noise, multiple-people and shape changes to facial components.
- To design develop, implement and test an efficient approach to automatic speaker recognition that can handle practical challenges faced by the state-of-the-art algorithms.

- To investigate the state-of-art in multi-focus and multi-exposure image fusion. In particular to investigate the robustness of algorithms in the presence of camera shake and investigate artefacts that may become visible when displayed in large screens.
- To design develop, implement and test an efficient approach to multi-exposure and multi-focus image fusion that can handle practical challenges faced by the state-of-the-art algorithms.
- To investigate the state-of-art in Vehicle Make & Model Recognition. In particular investigate the robustness of current algorithms in performing VMMR on CCTV video, which are often poor quality and not configured for VMMR tasks.
- To design develop, implement and test an efficient VMMR algorithm that can work on real CCTV footage captured under non-controlled conditions, is of poor-medium quality and not configured specifically for VMMR.
- To identify the limitations of the proposed algorithms and outline the future directions of research.

## 1.4 Contributions of Research

The work carried out in meeting the above aim and objectives of research has resulted in a number of contributions and novel ideas that extend the current state-of-art. These are presented in detail in Chapters 4-6. Majority of these research

results have already being published as conference proceedings and submitted as a journal paper. [Note: The full details of associated conference papers and submitted journal manuscript can be found in Appendix A]

### 1.4.1 Speaker Identification

In chapter 4 a novel approach based on CPD (Coherent Point Drift) is proposed for speaker detection in teleconferencing application based on the detection of lip movements. CPD is used to provide a measure of the correspondence distance between a lip identified in a video frame and a corresponding template lip that is known to be stationary. Moving lips are classified as lips having correspondence distances above an established threshold. A unique novelty of this work is the CPD algorithm's ability to analyse non-rigid transformations between point sets in a typical case of lip movement detection. It is noted that talking results in non-uniform deformations of the lips. The lip boundary is represented using edge points and prior to matching with the edge points of the template lip outliers and missing points are eliminated by the CPD approach. This increases the accuracy of the application of the calculation of the correspondence distance (as point correspondences are preserved despite the non-uniform deformations.), increasing the overall accuracy of lip movement detection. Our investigations revealed that in low illumination condition this novel approach resolves a major practical problem as edge detection of a lip will result in many missing points.

The accuracy of facial detection is improved by the use of temporal processing. Although faces are detected on a frame by frame basis, a feature tracking algorithm is used to track face like regions across the video frames before a decision is made about the reliability of the presence of a face. Once a face is identified as being reliable, the lip detection and movement analysis algorithms are applied.

Experiments revealed that the proposed algorithms can perform real time and when tested of a set of test video conferencing samples provided a 95.83% accuracy. The above work has been published as a conference paper [93]. Further the use of the above approach has been demonstrated to the users of UK Access Grid, receiving excellent feedback.

### 1.4.2 Multi Exposure/Focus Image Fusion with Camera Shake Compensation

Chapter 5 proposes a novel Multi-Exposure/Focus Image fusion algorithm which has the unique capability of compensating for camera shake prior to image registration. It is noted in all state-of-the-art algorithms and in the most popular mobile handsets (iPhone 4) the HDR (High Dynamic Range) algorithms proposed and implemented fail in the presence of camera shake, producing smudgy images. The proposed approach uses CPD (Coherent Point Drift) to register the multiple exposure images prior to fusion. CPD is applied to SIFT feature points detected on the two images being fused. To remove missing points and outliers effectively, RANSAC (RANdom SAample Consensus) algorithm is used, increasing the

accuracy of using CPD. In addition to the above novelty, in ensuring that images are registered by compensating for camera shake, a further novel approach is used to fuse the images. The fusion is performed in Contourlet transform domain with an additional stage of alpha blending and filtering to remove artefact formation. Experimental results have been produced to compare the subjective image quality achieved by the proposed approach with that of state-of-the-art algorithms including the top commercial products and it has been shown that the proposed approach outperforms its competitors.

The result of the above work has been presented in two different conference papers and has been extended and submitted as a journal paper [94, 95]. Further the approach is currently being considered for commercial exploitation by the collaborating industrial partner, Apical Ltd.

### 1.4.3 Vehicle Make and Model Recognition

Chapter 6 proposes a novel approach towards VMMR (Vehicle Make and Model Recognition) on video footage captured by CCTV cameras. In particular a novelty of this work is the proposed algorithm's ability to operate on video footage captured by CCTV cameras which are often of poor quality and have not been specifically configured for VMMR. The use of CPD allows the de-skewing of detected vehicles, thereby increasing the accuracy of recognition. Further the approach proposed includes temporal processing whereby the accuracy of recognition is improved by analysing multiple frames before decision is made on the vehicle make and model.

The literature review conducted revealed that there is no current system that has been demonstrated to work on general purpose CCTV footage. The proposed system achieved a classification rate of 95.83% on real CCTV footage.

A paper out of above work will submitted to the 18th International Conference on Digital Signal Processing 2013 that will be held in Greece in February 2013.

## 1.5 Organization of Thesis

This thesis has been organized into seven chapters as summarized below.

**Chapter 1** provides an overview of the application areas that the thesis is focused on. It further provides a brief summary of the research problems addressed in the thesis and the aim and the objectives of the research conducted. The chapter concludes with an insight to the original contributions made and an overview to the presentation structure adopted by the thesis.

**Chapter 2** presents details of the current state-of-the art in the three application areas this thesis is focused on. Where appropriate it also provides the state-of-art in the tools and other related applications widely used in the three application areas.

**Chapter 3** concentrates on providing the background knowledge related to the research context in which the novel techniques have been proposed. It covers the theoretical, conceptual and mathematical definitions utilized in chapters 4,5 and 6 in presenting the foundations of the original contributions of this thesis.

**Chapters 4, 5,** and **6** present the original contributions of this thesis, where Coherent Point Drift (CPD) has been used to provide solutions to a number of

practical problems in, speaker identification in video conferencing, multi-exposure/focus image fusion and vehicle make and model recognition. These chapters include specifically the novel methodologies of the approaches adopted, experimental setup used, design details, results and analysis of experiments carried out and conclusions reached.

**Chapter 7** concludes the research presented in this thesis with an insight into the future directions of research and possible enhancements to the proposed algorithms.

# CHAPTER 2

# Literature Review

## 2.1 Introduction

This chapter provides an insight to existing research on Speaker Identification, Multi Exposure and Multi Focus Image Fusion, Camera Shake Compensation and Vehicle MMR. These topics cover the contributory subject areas of this thesis, presented in chapter 4, 5 and 6.

The chapter is divided into several sections for clarity of presentation. Section 2.2 provides a review of literature on speaker identification. Section 2.3 provides a detailed literature review on Image enhancement techniques. Section 2.4 provides a review of literature on vehicle make and model recognition (MMR). Finally section 2.5 summarises, making conclusions and identifying open research problems in the selected areas of research.

## 2.2 Speaker Identification

With significant improvement to Internet technology over the past decade, teleconferencing has become a more useful choice for long distance meetings. Today teleconferencing can offer voice and video at the same time. One of the

complications in teleconferencing is, ‘speaker detection’, amongst several participants. Therefore it will be useful to offer users a close view of the current speaker and a system that automatically locates and tracks the speaker/s in such a system. Speaker detection is normally done with detecting the presence of human faces (i.e. face detection) and then observing which faces appear to belong to someone who is talking (i.e., consists of moving lips). The following sections present existing research on face detection, lip segmentation and movement detection and integrated systems which present speaker detection algorithms.

### 2.2.1 Face Detection

Face detection, although trivial under controlled conditions, is a challenging problem to solve, particularly in cases where the lighting, environment, face position and background are changeable. Face recognition takes face detection a further stage forward by being able to authenticate a detected face, belonging to an individual. In all face recognition algorithms, face detection is the first stage. This section reviews a number of face detection and recognition (noting that the initial stage is a face detector) approaches that have provided the basic motivation to the initial stage of the proposed speaker identification system.

M. Tistarelli et al. [1]

This work proposes a system, which uses two dynamic tracking cameras to fixate the face of the subject and select space-variant features from the appropriate facial

images. Features are extracted using a morphological filtering for a rough localization and an adaptive template matching. Then a matching algorithm based on a space variant representation of facial features is applied for identity verification. For identity verification a technique based on the Principal Component Analysis is used and subsequently the results of the verifications are compared. For testing the FERET data base is used. The reported equal error rate on overall performance is about 9%.

P. Viola et al. [2]

In this paper a new method is presented for face recognition which learns a face similarity measure from examples of given image pairs. The features computed are named as ‘rectangular’ features, which are proved to provide efficient representations. Feature selection is done afterwards using an AdaBoost algorithm, which is trained upon the face similarity function. The weighting procedure in AdaBoost is used to perform feature selection and subsequent classification. Experiments have been performed on the FERET facial database to prove that reducing the feature set can result in better recognition accuracy. 5.5% equal error rate is reported.

P. Yang et al. [3]

This paper focuses on reducing the dimensionality of feature set and the selection of the most discriminant features out of a large number of high dimensional Gabor features. Thus, a face recognition system based on AdaBoosted Gabor features [4, 5] has been proposed in this paper. The main contributions of the paper are: (1)

AdaBoost is successfully applied to face recognition by introducing the intra-face and extra-face difference space in the Gabor feature space; (2) An appropriate re-sampling scheme is adopted to deal with the imbalance between the amount of the positive samples and that of the negative samples. The experiments have been performed on the FERET database. The results demonstrate the effectiveness of algorithm when reducing and selecting the number of features. With 700 selected Gabor features, 95.2% accuracy rate is achieved.

M. Bicego et al. [6]

This paper investigates the application of the SIFT approach proposed originally by [59] in the context of face authentication. In order to determine the real potential and applicability of the method, different matching schemes have been proposed and evaluated. The matching schemes between the test and training set adopted are different to the original SIFT matching scheme and includes; minimum pair distance, matching eyes and mouth, and matching on a regular grid. The average Equal Error Rates for a pair of distance is 12.92%, for eye and mouth matching is 10.88% and for a regular grid is 7.58%. The system has been tested on the BANCA database.

B. Heisele et al. [7]

In this paper a component-based method and two global methods for face recognition is presented. The focus of this paper is to evaluate the system performance against varying pose. In the component-based system, facial components are located, extracted and combined into a single feature vector which

is classified by a Support Vector Machine (SVM) [8]. The two global systems recognize faces by classifying a single feature vector consisting of the grey values of the whole face image. In the first global system a single SVM classifier is trained for each person in the database. The second system consists of sets of viewpoint-specific SVM classifiers and involves clustering during training. The tests have been performed on a database which includes faces rotated up to about 40 degrees in depth. Results indicate that the overall component-based system outperforms both global systems for recognition rates larger than 60%.

### 2.2.2 Lip Segmentation and Lip Movement Detection

Segmenting lip area and then lip movement detection are the next steps in a speaker identification system. This section reviews a number of existing lip segmentation and lip movement detection approaches.

J. Luettin et al. [9]

This work describes a lip reading system based on visual features extracted from grey level image sequences of the speaker's lips. For tracking the lip contours dynamic shape models are used while visual speech information is taken out from the contour shape. The geometrically analysed lip features include information for the shape and intensity information of the lips. This approach uses Hidden Markov Models (HMMs). Speaker independent word recognition is used besides lip reading

as a final stage. The proposed work is tested on a sub-set of the Tulips 1 database and an accuracy rate of 72.2% is reported.

P. Dalka et al. [10]

In this paper an algorithm for lip movement tracking and lip gesture recognition based on artificial neural networks is proposed. It uses a precise lip shape attained by means of a lip image segmentation algorithm that uses fuzzy clustering. The experiments are conducted on 102 collected face images obtaining 90% recognition rate. However the number of recognized gestures is very limited in this work.

F. Haider et al. [11]

This paper investigated the use of lip movement for the purpose of speaker and voice activity detection using three techniques: neural network based, naive Bayes classifier based, and using Mahalanobis distance which is based on relations among variables by which different shapes can be identified. Two approaches were considered, one speaker dependent and the other speaker independent. In speech/silence detection, the experiments indicate 78.31% accuracy, when using lip movement along with the named techniques. The results indicate that the neural network classifier based approach produced more accurate results than the other two techniques in the speaker dependent approach. However, when the speaker independent approach is used, the results indicate that the naive Bayes classifier performs best with an accuracy of 64.56%.

M. Bendris et al. [12]

In this paper a method evaluating pixel disorder degree directions around the lip was proposed for movement detection. Visual information in a Television context has been used to find the relationship between speech sequences and corresponding faces. The assessment has been conducted on a large database of TV-shows. The proposed method has achieved an equal error rate of 17.5%, which is relatively better than the method based only on the difference of the pixels.

S. Siatras et al. [13]

This paper focuses on using only visual information. The key feature in this method is the application of signal detection algorithms to extract features from the mouth region. A statistical algorithm that utilizes the average value and standard deviation of the number of pixels with luminance information on the mouth region of a speaking person for the effectual description of visual speech and silence in video sequences has been derived. For a multi-person environment the lip activity detection method is used to determine the active speaker. The proposed algorithm reports a 1.16 equal error rate but does not provide satisfactory results in some conditions like poor lighting, non-frontal faces or faces distant from the camera.

### 2.2.3 Speaker Detection

Creating a Speaker Identification system is the core idea of the first part of thesis. In this section several speaker identification methods are presented to obtain a basic idea of previously literature.

T. Hazen [14]

This work combines the application of existing face and speaker identification techniques for user verification on a handheld device to demonstrate a fused multi-modal system. For face detection a fast hierarchical classifier is used to localize the face in the captured images and then a component based face detector is applied to correctly localize the face and detect facial components. To identify speaker, a speaker dependant speech recognizer has been trained and used. The test was performed on a small size of data set. A 90% reduction in user verification equal error rate is reported using the proposed system.

D. Shiell et al., [15]

This paper has described an automatic system for person identification, by detecting, tracking and identifying a speaker by extracting visual features from the speaker's mouth area. An Active Appearance Model (AAM) is used for face and mouth tracking on the extracted face area. Then region-of-interest (ROI) around the mouth is extracted for visual features and the 2D discrete cosine transform (DCT) is applied. The experiments indicate that the proposed system achieves an accuracy

rate of 59.3% compared to using the ground tracking data, which achieves a success rate of 52.3%.

H B Kekre et al. [16]

This paper proposes Vector Quantization for speaker identification. Vector Quantization is used to extract features based on the lossy compression. Codebooks are produced from the speech samples for training and testing. This approach is completely based on analysing the audio signal only and is hence outside the scope of the work carried out in this thesis.

D. Pullella et al. [17]

This paper studies the combination of missing data processing and feature selection to improve a speaker identification system where some data is missing. The development of a subset of classified filter-bank features allows the modification of binary Time Frequency (TF) reliability masks by the elimination of unreliable presence errors for non-discriminative bands. Experimental results indicate that the proposed method can perform well in bottom-up approach under presence of noise. However if the reliability mask is incorrectly estimated the combined system considerably underperforms the traditional bottom-up only approach.

J. C. Wang et al. [18]

This paper proposes an approach for speaker identification based on a subspace-based enhancement technique and probabilistic support vector machines (SVMs). First, a perceptual filter bank is created from a psycho-acoustic model into which

the subspace-based enhancement technique is incorporated. For subspace-based enhancement technique a filter bank is formed from a psycho-acoustic model and SNR of each sub-band is calculated to efficiently defeat environmental background noises in speech. Then, probabilistic SVMs which are generated from distance ratios classify and validate the speaker from the enhanced speech. For experiments 20 speakers from AURORA-2 database with added background noise have been tested. The SVM classifier has a rate of success of about 90.1% when used for speaker identification. This method underperforms in the presence of background noise. However by applying the speech enhancement, the recognition rate is improved from 20.3% to 48.6% in speaker identification.

S. Kwoon et al. [19]

In this paper feature vectors are used for speaker identification. Feature vectors contribute to perception. Speaker model is trained in a way to select only suitable feature vectors to overcome conclusion errors. For testing a 400-speaker data subset has been gathered from the Speaker Recognition Benchmark NIST Speech corpus. The experimental results indicate that this approach improves overlapped speaker model situation in a given feature space. The authors claim to achieve an accuracy of 85% with one fourth the length of utterances using the proposed method.

## 2.3 Multi-Exposure and Multi-Focus Fusion

Creating High Dynamic Range (HDR) images using Standard Dynamic Range (SDR) images captured using traditional sensors requires an important stage, i.e., fusion of multi-exposure images. Further forming an image with a high depth of view requires the fusion of multi-focus images. In many cases multi-exposure image fusion and multi-focus image fusion can be achieved using the same image processing algorithms. This section presents the existing multi-exposure and multi-focus image fusion algorithms.

J. An et al., [20]

This paper proposes a multi-exposure image fusion algorithm for HDR imaging. Image fusion is performed by calculating the weighted sum of multi-exposure images, which takes contrast, saturation and well-exposedness into account. As for the removal of the ghosting effect which is an imitation of the transmitted image, offset in position, the weighted values of pixels are maintained within upper bounds. However it has been shown that this method is not capable of totally removing the ghosting effect. After zooming on the generated image, some blocking artefacts and different illumination levels are visible.

I. Zafar et al., [21]

The algorithm proposed in this work is for both multi-focus and multi-exposure image fusion and is based on Discrete Cosine Transform (DCT). It is shown that

fusion in DCT domain accelerates effective fusion without losing image data. The experimental results are compared with some commercial software packages and the performance of the approach proposed is claimed to be visually better.

K. Kotwal et al., [22]

The technique proposed for multi-exposure image fusion is based on Euler-Lagrange method which inserts the preferred properties of the output image into an objective function, and delivers an iterative resolution to reach the desired result. This method operates on the data without any information from camera response function or exposure settings. The experimental results are visually compared to [20] and have produced better contrast.

S. Lee et al., [23]

In the multi-exposure image fusion algorithm presented in this paper similarity measures between multi-exposed images are used for motion estimation to provide a solution to camera movement. A statistical measure of mutual information is used to improve the estimation of brightness gaps between different exposures. Experiments were conducted on 100 pairs of images and were shown to have a success rate of 86%.

M. Qiguang et al., [24]

This research proposes an image fusion method based on Contourlet transform. It is argued that Wavelet transforms are unable to represent non-linear shapes/curves in images effectively. The use of Contourlet transform is proposed as a solution for

image fusion. The images are decomposed using a Laplacian Pyramid. Then a directional filter is applied to identify the direction of information representation of the image effectively. As the fusion rule for low frequency sub-bands, averaging of coefficients is used and for high frequency sub-bands the maximum of the absolute frequency is used. The experimental results are compared to Laplacian, Ratio pyramid, Gradients pyramid and Shift Invariant Discrete Wavelet Transform (SIDWT) methods and produced improved results.

A. Goshtasby [25]

This technique for image fusion divides the image domain into homogeneous blocks and for each block selects the image that includes the most amount of data inside that block. The particular images are then merged jointly using monotonically decreasing merging functions that are cantered at the blocks and have a sum of 1 everywhere in the image domain. The optimal block size and width of the blending functions are determined using a gradient-ascent algorithm to maximize information content in the fused image. Entropy is used as the measure for optimization when fusing the images.

M. Block et al., [26]

This method is for fusing images of multi-exposure text-documents. OCR-recognition rate is used as a performance comparison benchmark. This algorithm merges the high-pass filtered images instead of original images to increase the recognition rate. For the experiments, 3 exposure set of 40 different documents with several fonts and font-sizes images were used. The accuracy of the recognition is

determined according to the Needleman-Wunsch algorithm and a recognition rate of 0.95 is reported.

T. Mertens et al., [27]

In this method multiple exposure images are fused by using quality measures including saturation, contrast and level of exposure as defined by the authors. The processing is carried out in a multi-resolution manner by calculating a weighted average along each pixel, across the multiple images, using weights calculated from quality measures. In the work proposed the method has not being tested in presence of noise. The method is visually compared with tone mapping techniques such as of Durand, Reinhard and Li.

L. Yang et al., [28]

In this work an algorithm for multi-focus image fusion based on Contourlet transform, region variance and local energy is proposed. The multi-focus images are first decomposed using Contourlet transform to separate the characteristics of multi-scale, localization, directionality and anisotropy. Experimental results indicate the proposed method can extract image features more efficiently and achieve improved fusion performance compared with the gradient pyramid algorithms and Wavelet based fusion algorithms, with an increment in mean square error of 34.8% and 42.6% respectively.

S. Li et al., [29]

An algorithm for multi-focus image fusion using Wavelet and Curvelet transforms is proposed in this proposed work. First each image is decomposed using Curvelet transform, followed by a fusion process for all coefficients carried out using Wavelet-based image fusion. Final image reconstruction is done by performing the inverse Curvelet transform. The results are compared to a Wavelet method and proven to perform better.

L. Ding et al., [30]

In this method Contourlet Transforms that provides multi-scale, localization, directionality and anisotropy advantages is used for multi-focus image fusion, using an initial Wavelet based Contourlet decomposition (WBCT). The images are decomposed to low and high frequency images utilizing the Wavelet based Contourlet transform. In the low-frequency image the fusion rule is based on regional variance and in high-frequency Contourlet sub-bands the rule is based on the local energy. Compared to the traditional gradient pyramid algorithm and Wavelet fusion algorithm, it is shown that the proposed algorithm is capable of improved performance. Mean Square Error (MSE) is used as a comparison measure. Amongst Pyramid, Wavelet, and Contourlet fusion methods, the proposed method has the smallest MSE.

L. Tang et al., [31]

A new image fusion scheme based on the WBCT is presented. Firstly, the WBCT is used to perform a multi-scale and multi-direction decomposition of each image.

Then the WBCT coefficients of fused image are constructed using multiple calculations according to different fusion rules. The experimental results show that this fusion scheme is effective and the fused images are better than that obtained from the Laplacian pyramid transform, the Wavelet transform and the Contourlet transform. The sharpness and the entropy is the biggest, the mean cross entropy and the root cross entropy is the smallest, using proposed method. The fused result using WBCT-based method is the best amongst the tested methods.

M. Choi et al., [32]

The approach presented in this paper is based on the Curvelet transform and aims to enhance edges in the processing. In order to enhance spatial declaration the edges are improved. Curvelet-based image fusion technique presents more information and detail in the spatial and spectral domains simultaneously. The IKONOS image database is employed to demonstrate the best possible fusion result. Relatively better results are achieved compared to Wavelet transform based fusion and HIS (Intensity, Hue and Saturation) based on combination entropy, mean gradient and correlation coefficients measurements.

F. Sroubek et al., [33]

In this paper an algorithm for multi-focus image fusion based on the application of space-variant and edge-oriented window neighbourhoods is proposed. This approach helps to obtain a more accurate decision map, especially in the presence of outliers which can create unrealistic peaks. Based on ground truth the percentage of incorrect decisions (PID) is calculated to evaluate the quality of decision map and

percentage mean squared error (PMSE) is calculated to evaluate the quality of the fused image.

T. Zaveri et al. [34]

This paper a two-step structure for image fusion is proposed. In the first step the given source images, which are to be fused, are segmented. The image segmentation process is assumed to be a graph partitioning problem. A novel global criterion, ‘normalized cut’, is used for segmenting the graph. The output of the segmentation step is the main contribution of the proposed method. It is shown that a small alteration in the segmentation outcome can make a significant difference to the final result. Subsequently the second and last step is to create the fused image using fusion rules. The experimental results are based on root mean squared error (RMSE) and mutual information (MI) measurements are compared to DWT, Contourlet and region-based techniques and indicate that the proposed approach produces more detail compared to direct use of individual pixels for fusion.

## 2.4 Image Registration

When images are to be fused it is essential to check whether they are registered with each other. Failing to do so will lead to smudgy fused images that are useless for any purpose. In this thesis we provide a novel approach to image registration. This section presents the current state of the art in image registration.

G. Ward [35]

The work proposed in this paper uses thresholded bitmaps to accelerate image operations. The image pyramid of each exposure is registered with another by using inexpensive shift and difference operators. The reported success rate of this method is about 84%. Unsatisfactory results were obtained due to image rotation. The proposed algorithm cannot detect misalignments and lacks rotation correction.

T. Grosch [36]

In this work Median Threshold Bitmaps are used for image registration and to find the best translation and rotation angle alignments. This work is an improved version of [35]. By testing a predicted colour for each pixel image ghosting areas can be detected and filled, with a region similar to that from the reference image. The performance of algorithm was tested on 180 image sequences obtained from hand-held cameras. The proposed algorithm needs manual post-processing to correct additional shortcomings such as de-blurring and correcting misalignments in fused images. The results are compared with Photoshop CS2 and in the majority of cases, a visually better image quality is observed with proposed method.

A. Tomaszewska et al., [37]

In this work, SIFT is used to identify the keypoints in the multiple-exposure images and the keypoints are subsequently matched for registering the images by calculating the transformation matrix using the direct linear transform technique. For image fusion, the algorithm proposed by M.D. Grossberg, et al [38] is used which is based on polynomial approximation of camera response function. Using

this function the correct radiance pixel value is calculated to fuse images together. The foremost outcome of the proposed algorithm is an image registration method which can be used to eliminate misalignments between images in a set of different exposure images. The proposed algorithm has a disadvantage that it is limited to using only 4 feature points after RANSAC is applied. At low levels of illumination this could lead to confusing matches due to ambiguity of feature points.

A. Myronenko et al. [39]

In this paper a probabilistic method for non-rigid registration of point sets is proposed. The registration is assumed as a Maximum Likelihood (ML) estimation problem with motion consistency limitation over the velocity area in a way that one point set moves coherently to align with the second set. Also the Expectation Maximization (EM) algorithm is derived for the ML optimization with deterministic strength. The CPD method concurrently finds the non-rigid transformation and the correspondence between two point sets without making any prior assumption of the transformation model except that of motion coherence. This method can estimate complex non-linear non-rigid transformations, and is shown to be accurate on 2D and 3D examples and robust in the existence of outliers and missing points.

## 2.5 Vehicle Make and Model Recognition

The classification of vehicles has been a focus of interest in the past, for traffic control systems and toll levy automation. Most of these systems considered, attempt

to classify the vehicles into broad categories e.g. cars, heavy goods vehicle etc. In this thesis we take this a further step by proposing an algorithm that can recognise the make and model of a vehicle. This additional attribute will help improve the level of security offered by the traffic control system. In this section we present existing work on vehicle make and model recognition.

X. Clady et al., [40]

This paper proposes a VMMR system based on oriented contour points. A ROI (frontal view) is extracted using the license plate as a reference and oriented contour points were extracted from the ROI and used as features. A voting system is developed that is used to classify the test vehicle image into a make and model class. Experimental results claim 93.1% recognition rate on a dataset consisting of 50 classes of vehicles.

F. M. Kazemi et al., [41]

In this paper a comparative study of the performance of Fourier, Wavelet, and Curvelet transform features for VMMR is presented. In the experiments, the transforms are applied on the region of interest and the coefficients of the transforms mentioned above are fed to a k-nearest neighbour classifier. The best recognition rates are achieved using all the Curvelet transform coefficients as features. It should be noted that in this experiment only 5 classes of vehicles are examined, and as such results may vary with a larger number of classes.

F. M. Kazemi et al., [42]

In this work Curvelet transform features with a Support Vector Machine (SVM) is used as an alternative solution. After selecting a ROI from vehicle's rear view, Curvelet transform coefficients are extracted at different scales and directions. The standard deviations of the coefficients for each scale and direction are used as features, instead of the coefficients of the transform; this is done to reduce the dimensionality of features. The experiments are done with a k-nearest neighbour classifier and two SVM classifiers - One versus One and One versus All. An excellent result of 99% is obtained by selecting the most useful scales from the transform and using the One versus One SVM classifier. The dataset used is same as [41].

S. Rahati et al., [43]

This paper proposes a solution similar to the one used by [42], but replaced the Curvelet transform features with Contourlet transform features. Contourlet transform coefficients at different sub-bands and directions are computed. The standard deviations of Contourlet transform coefficients from selected scales with the most useful information are fed to the SVM classifier. A 99% recognition rate is reported using the dataset in [41].

I. Zafar et al., [44]

This work proposes an improvement to the technique proposed by [43]. The use of a localized Contourlet feature extraction technique is proposed as opposed to the use of standard deviations of the Contourlet coefficients. A SVM is used for

classification. An accuracy rate of 94% is obtained using 10 samples per class to train the SVM. Further an accuracy of 52% is reported when the technique in [43] is used. This technique significantly improves the approach proposed in [43].

M. M. Arzani et al., [45]

In this work the use of combined Wavelet and Contourlet features for VMMR is proposed. Features from half of the frontal view image of the vehicle are extracted as opposed to other systems that used the full frontal view image. The best recognition rate of 97.35% is achieved by taking the best features from the Wavelet decomposition and combining them with the best features from the Contourlet decomposition, these features are fed into a SVM for classification.

S. M. Sarfraz et al., [46]

In this paper the use of Local Energy Based Shape Histogram (LESH) features is proposed. LESH features are extracted from a ROI taken from the front view of the vehicle. The extracted features are modelled in a similarity feature space using a probabilistic Bayesian framework. Using Bayes rule, the posterior over possible matches is computed and the highest score is selected as the make and model class. A high recognition rate of 94% is obtained in the reported experiments. This approach has the advantage that only a single image is required as a reference image after offline training.

S. M. Sarfraz et al., [47]

The technique proposes in this paper is to solve the VMMR problem that does not require vehicle segmentation, but detects salient regions called “Patches”. A local description for each patch is extracted using Local Energy Shape Histogram. The local description is modelled in a similarity feature space classified using a probabilistic Bayesian framework. 94% accuracy is achieved using the same dataset used in [48] and 62% accuracy on a new vehicle dataset collected in uncontrolled conditions.

A. Psyllos et al., [48]

A hierarchical technique for solving the VMMR problem is proposed in this research paper. The make recognition and model recognition are separated. A ROI is extracted from the vehicle’s front view using the location of the license plate as reference; from the ROI the manufacturers logo is extracted and then the features extracted from the logo are used to classify the unknown vehicle into a particular make (e.g. ford, Renault etc.) using a Probabilistic Neural Network (PNN). After which, Scale Invariant Feature Transform (SIFT) features are extracted from the vehicle’s image which are used to determine the model also using a PNN. A 59% success rate is obtained with proposed system. Further a fast colour recognition module which gave a 90% recognition rate is introduced.

B. Daya et al., [49]

This research provides a VMMR system that used geometric parameters to represent the different makes and models of vehicles. Three geometric

measurements were taken from the ROI selected from the vehicle's frontal view. These geometric measures are normalized using the height of the licence plate, after which they are passed to a Neural Network for classification. A recognition rate of 97% is obtained from a dataset of 12 models. The challenge with this approach is that it will be unable to properly classify vehicles that have other differences other than size.

J. Prokaj et al., [50]

This approach attempts to recognize the vehicle's make and model from a video clip taken at an arbitrary viewpoint. Using prior knowledge of the ground plane to impose constraints on the virtual camera motion, the vehicle pose from each frame of the video is estimated and its 3D motion on the plane is calculated using a structure from motion algorithm. The 3D models of vehicles in the database are rotated to the same pose as the calculated 3D structure; then features similar to SIFT features are extracted but with rotation invariance disabled. Using these features, a video-model similarity metric is computed and the model with the highest score is selected as the make and model. The system is tested with 20 video clips and a database with 36 vehicle models and reports a 50% correct classification rate.

## 2.6 Summary and Conclusion

In this chapter, existing algorithms in speaker identification, image fusion and VMMR have been introduced and critically analysed. In addition, key algorithms in

each area and relevant literature have been reviewed. From the literature review, as documented it is obvious that there are several research gaps that can benefit from further research.

The proposed solutions for Speaker Identification in videoconferencing are based on speech identification or visual information processing which lack accuracy and can come across important challenges when faced with addressing practical challenges in a usual videoconferencing.

The literature in the area of multi-exposure and multi-focus image fusion for creating HDR images reveals that the existing algorithms are unsuccessful in the presence of camera shake which is a serious inadequacy for application in hand held devices. Further demonstrating the fused images on high resolution, results in numerous artefacts for instance, ‘image ghosting’, ‘blockiness’, etc.

All existing approaches on Vehicle Make & Model Recognition (VMMR) have been applied under controlled conditions, assuming that the cameras are precisely set up and are rigidly fixed and situated. However there is a need to categorise the vehicle make and model based on general purpose CCTV video footage, which are average to poor quality and recognize approaching vehicles at an angle that makes vehicles appear skewed in shape. These create additional challenges to the VMMR algorithms which existing techniques cannot satisfy these challenges.

Considering the shortcomings of the existing algorithms, a number of novel approaches to speaker identification, image fusion and VMMR have been proposed in chapters 4-6. The research motivation behind each approach has been presented in the relevant chapters. Chapter 3 introduces the reader to the fundamental concepts/theories used within the above contributory chapters.

# CHAPTER 3

# Background

## 3.1 Introduction

The work presented in this thesis is based on a number of fundamental concepts and theories in different areas of mathematics, computing and engineering. This chapter introduces the readers to the above, supported well by mathematical notations and definitions.

## 3.2 Energy

Generally, energy in signal processing resembles the mean squared value of the signal.

Given an image $I$, suppose $f_{ROI}\ (x, y)$ (where $ROI$ is the region of interest) is the coefficient in position $(x, y)$; energy of $ROI$ can be calculated as:

$$E_{ROI}^{I} = \sum_{x,y \epsilon ROI} f_{ROI}^{I}\ (x, y)^2 \qquad 3.1$$

## 3.3 Coherent Point Drift

One important requirement in the registration of digital images is to find important correspondences amongst two point sets to improve the core transformation that maps the first point set to the second. Point set registration is crucial for numerous computer vision applications. The associated transformations are typically divided into two groups: rigid and non-rigid.

### 3.3.1 Rigid transformations

Having two point sets from two different sets of data, a rigid transform retains distances amongst each and every pair of points. The Rigid transform includes reflection, translation, rotation, and/or their combination. All objects retain the identical outline and size after a correct rigid transformation.

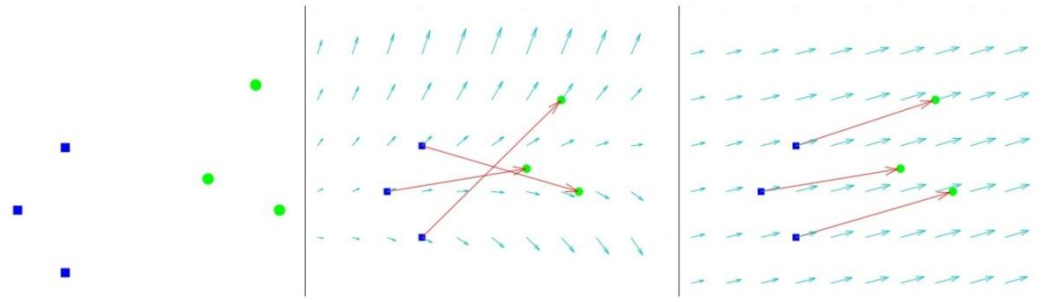
### 3.3.2 Non-Rigid

A 'Non-rigid' or 'elastic' transform allows nonlinear transforms and curved transforms. These transforms are accomplished by warping the chosen image to line up with the reference image. Affine is the simplest non-rigid transform which is also capable of scaling, rotation and skew. Non-rigid transforms occur in numerous real-world problems such as deformed motion tracking, shape recognition, and medical image registration. A very accurate non-rigid transformation model is often

unknown for a point-set and is challenging to model. Due to the significant amount of transformation factors, the non-rigid point set registration methods are more sensitive to noise and outliers.

The concept of motion coherence was proposed in the Motion Coherence Theory [53]. Coherent Point Drift (CPD) [54], is a probabilistic technique for non-rigid registration of point sets. The registration is assumed as a Maximum Likelihood (ML) [51, 52] approximation problem. Gaussian mixture model (GMM) [76] centroids from first point set is fitted to the second point set by maximizing likelihood. Subsequently GMM centroids are moved coherently as a whole to reserve the topological structure of the point sets (Figure 3.1). The motion coherence limitation is formulated and a solution of normalized ML approximation is developed through the deviation method, which results in a well-designed kernel model [51, 52]. Finally the Expectation Maximization (EM) [51, 52] algorithm for the penalized ML enhancement with deterministic annealing is developed.

The CPD technique concurrently finds the non-rigid transformation and the correspondence amongst two point sets using motion coherence, regardless of the transformation model. This technique can estimate complex non-linear non-rigid transformations and is strong in the presence of outliers and misplaced points. Non-rigid registration assumes that the underlying transformation - required to align point sets - is complex, locally non-linear and the insight is that the points close to one another tend to move coherently.



(a) (b) (c)

Figure 3.1: (a) Two given point sets. (b) Coherent velocity field. (c) Velocity fields which are less coherent for the given correspondences. [39]

According to [54] algorithm of CPD can be summarized as below:

Given two point sets:

$X_{N\times D}$ - Reference point set (data points)

$Y_{N\times D}$ - Template point set (GMM centroids);

- Consider the points in $Y$ as the centroids of a Gaussian Mixture Model, and fit it to the data points $X$ by maximizing the likelihood function.
- Denote $Y_0$ as the initial centroid positions and define a continuous velocity function v for the template point set such that the current position of centroids is defined as

$$Y = v(Y_0) + Y_0 \qquad 3.2$$

.

- Find $Y$ by MAP. Minimize:

$$E(Y) = -\sum_{n=1}^{N} \log \sum_{m=1}^{M} e^{-\frac{1}{2}\left\|\frac{X_n - Y_m}{\sigma}\right\|^2} + \frac{\lambda}{2}\phi(v) \qquad 3.3$$

- $\phi(v)$ is the regularization to ensure the velocity field v (displacement) to be smooth. One choice is to measure the high frequency content:

$$\phi(v) = \int \frac{|\tilde{v}(s)|^2}{\tilde{G}(s)} ds \qquad 3.4$$

where $\tilde{v}$ indicates the Fourier transform of the velocity. $\tilde{G}$ represents a symmetric low-pass filter.

- It can be shown using a variational approach that the function which minimizes $E$ has the form of radial basis function:

$$v(z) = \sum_{m=1}^{M} w_m \, G(z - y_{0m}) \Rightarrow Y = Y_0 + GW \qquad 3.5$$

where $G_{M\times M}$: Gaussian affinity matrix

- The motivations to choose a Gaussian kernel form for $G$:
  - ✓ It satisfies the required properties (symmetric, positive definite, and $\tilde{G}$ approaches zero as $\| s \| \to \infty$).
  - ✓ Gaussian form in both frequency and time domain without fluctuations.
  - ✓ The flexibility to control the range of filtered frequencies and thus the amount of spatial smoothness.
  - ✓ It is equivalent to Motion Coherence Theory prior: sum of weighted squares of all order derivatives.

$$\int \sum_{m=1}^{\infty} \frac{\beta^{2m}}{m!\, 2^m} (D^m v)^2 \qquad 3.6$$

- EM optimization [51, 52]

Minimization of $E$ in Equation 3.3 is equivalent to minimization of its upper bound $Q$:

$$Q(W) = \sum_{n=1}^{N} \sum_{m=1}^{M} P^{old}(m|X_n) \frac{||X_n - Y_{0m} - G(m,.)W||^2}{2\sigma^2} + \frac{\lambda}{2} tr(W^T G W) \quad 3.15 \qquad 3.7$$

The minimum of $Q$ function is a solution of a linear system of equation:

$$[diag(p.1)G + \lambda\sigma^2 I].W = [PX - diag(P.1)Y_0](M - step) \qquad 3.8$$

where $P$ is the matrix of posterior probabilities with elements

$$p_{mn} = \frac{e^{-\frac{1}{2}\|\frac{X_n - Y_m}{\sigma}\|^2}}{e^{-\frac{1}{2}\|\frac{X_n - Y_m}{\sigma}\|^2} + (2\pi\sigma^2)^{\frac{D}{2}}/a} \qquad 3.9$$

- An additional uniform probability density function component is added to the mixture model in order to account for outliers.

CPD pseudo code according to [54] is:

1- Initialize parameters $\lambda, \beta, \sigma$

2- Construct $G$ matrix, initialize $Y = Y_0$

3- Deterministic annealing:

- EM optimization:

  E-step: Compute $P$

  M-step: solve for $W$ from Equation 3.7

Update $Y = Y_0 + GW$

- Anneal $\sigma = \alpha \times \sigma$

4- Compute the velocity field: $v(z) = G(z,.)W$

5- Find the correspondence from posterior probabilities $P$

An overview of how the algorithm works is shown in Figure 3.2, 3.3 and 3.4.

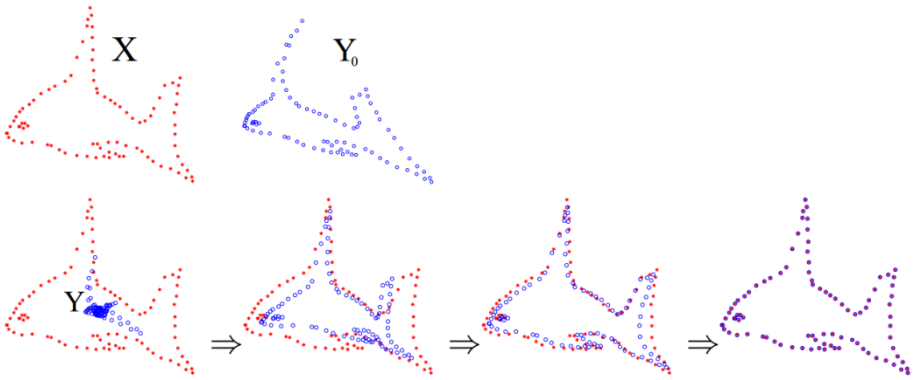



Figure 3.2: Shape alignment: blue point set are being registered to red point set. [39]

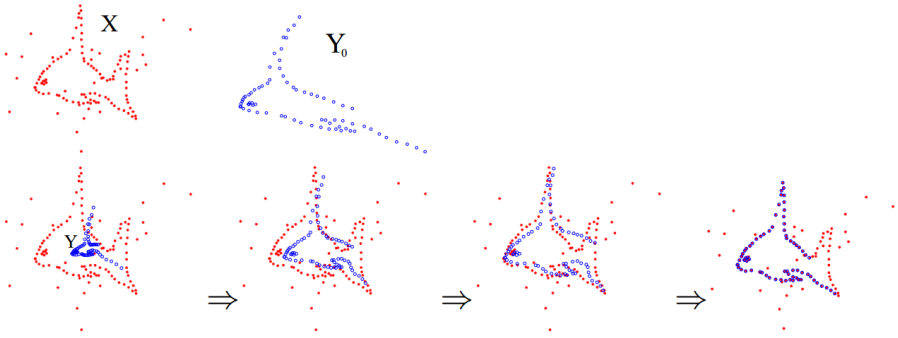



Figure 3.3: The red point set (Reference) is corrupted by noise and the blue point set (Template) has a missing tail to make the registration challenging. [39]

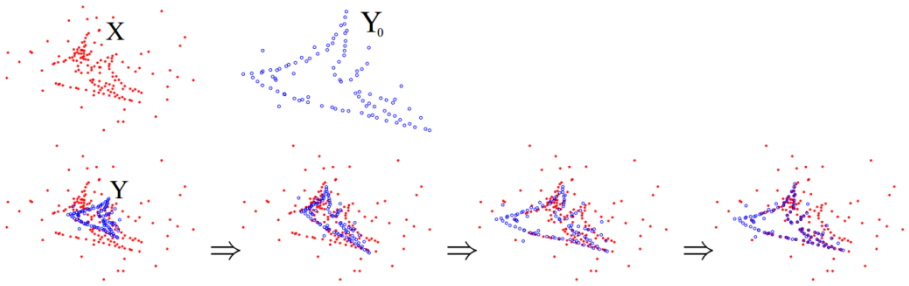



Figure 3.4: The red point set (reference) has a missing head and the blue point set (template) has a missing tail. Both sets are corrupted by noise. [39]

## 3.4 Scale Invariant Feature Transform (SIFT)

Scale Invariant Feature Transforms, first proposed by David Lowe [55] is an approach that has the ability to extract distinctive, invariant features from images. These features can be used for the reliable matching of images, thereby making SIFT highly suitable for robust image retrieval and object recognition applications. It has been shown that SIFT features are theoretically fully-invariant to scale and rotation and that they are partially invariant (i.e. robust) to affine distortion, noise, change in 3D viewpoint and illumination. A significant number of features can be extracted from typical images, which densely cover the image over the full range of scales and locations. For example, a typical image of size $500 \times 500$ pixels may give rise to about 2000 stable features (the exact number depends on both image content and choices of parameters). Further the number of feature points selected can be flexibly varied, by the suitable selection of parameter values. In addition, the SIFT features are highly distinctive, which allows a single feature to be correctly

matched with high probability against a large database of features, providing a solid basis for object identification/recognition [56].

### 3.4.1 Detection of Scale-Space Extrema

Scale-space construction: The first stage of computation involves the search for stable features (i.e. key points) over all scales and image locations (i.e. scale space of an image). The scale space of an image is defined as a function $L(x, y, \sigma)$, that is produced from the convolution of a variable-scale Gaussian, $G(x, y, \sigma)$ with an input image $I(x, y)$:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \qquad 3.10$$

Where * is a convolution operator

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{\frac{-(x^2+y^2)}{2\sigma^2}} \qquad 3.11$$

Stable key-point locations in scale space are found using extrema in the Difference-of-Gaussian (DoG) function convolved with the image, $D(x, y, \sigma)$, which can be computed from the difference of two nearby scales (filtered images) separated by constant multiplicative factor k:

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma) \qquad 3.12$$

Figure 3.5 illustrates the formation of a scale space for a given image. For each octave of scale space, the initial image is repeatedly convolved with Gaussians to produce the set of scale space images shown on the left side of Figure 3.5. Adjacent

Gaussian images are subtracted to produce the DoG images on the right. After each octave, the Gaussian image with σ twice that of the original is sub-sampled and used to construct the next octave.

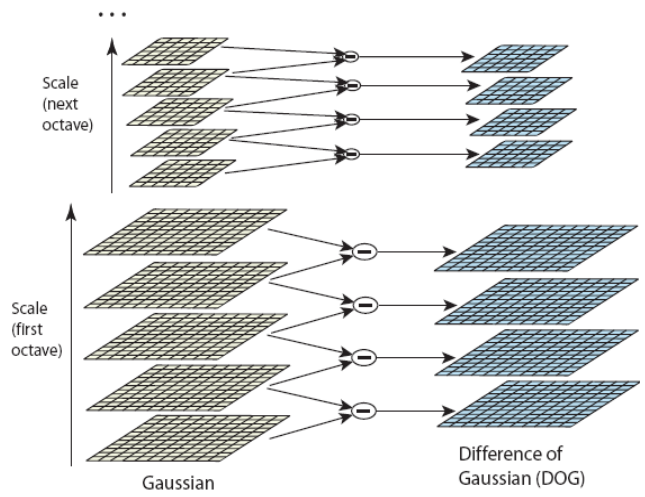



Figure 3.5: Gaussian blurred images at different scales and their respective DOG images. [55]

For the details on selection of number of images within each octave of scale-level and value of k readers are referred to [55].

Extrema Detection: An extrema is defined as any value in the DoG greater than or smaller than all its neighbours in scale-space. To find the extrema points each pixel is compared to 8 neighbours at the same scale plus 9 neighbours each from scales above and below (Figure 3.6). If the pixel is a local maximum or minimum, it is selected as a candidate key-point.

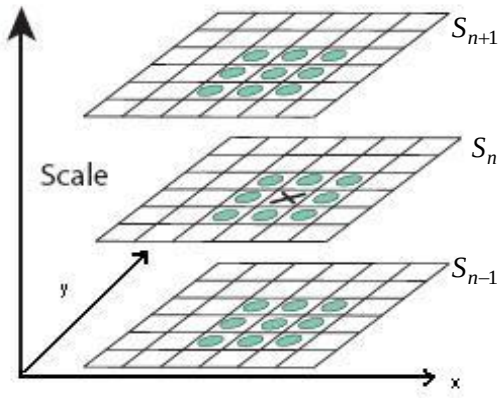



Figure 3.6: Local extrema detection, the pixel 'x' at $S_n$ is compared to total of 26 neighbouring pixels at same level and levels above and below. [55]

### 3.4.2 Key-Point Localization

The candidate key-points selected using the procedure outlined in section 3.4.1 are further refined (key-point localization) by measuring their stability. A detailed model is subsequently fit to pixels surrounding the candidate key-points, for determining location, scale and principal of curvature. By using this information, key-points having low contrast or are poorly localized along edges are rejected. The remaining key-points are classified as stable key-points and are used in subsequent analysis.

### 3.4.3 Orientation Assignment

Each of the stable key-points is then assigned with an orientation. The orientation of a key-point is calculated by computing a gradient orientation histogram for pixels in

the neighbourhood of the key-point as illustrated in Figure 3.7. The peaks in the histogram are considered to correspond to the dominant orientations. The maximum value and the 80% of the maximum value in histogram are used to create separate key-point for these directions.

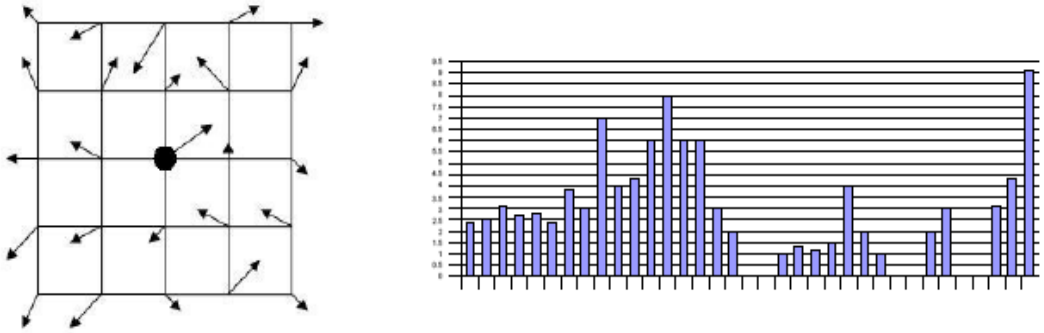


Figure 3.7: Left: Key-point (represented by a black dot) surrounded by neighbouring points. Arrows show the orientations of points. Right: Orientation histogram of magnitude sums for all the points within a window around the key-points. [55]

### 3.4.4 Key-Point Descriptor Computation

The orientation of the key-point is then used to find out the feature descriptor of the key-point. A descriptor is computed for the local image region that is as distinctive as possible at each key-point. The image gradient magnitudes and orientations are sampled around the key-point location. These values are illustrated with small arrows at each sample location on the left-hand image in Figure 3.8. The contribution of each pixel is weighted by the gradient magnitude, and by a Gaussian weighting function with $\sigma$ 1.5 times the scale of the keypoint. Each descriptor is a 128 (i.e. $8 \times 4 \times 4$) element feature vector.

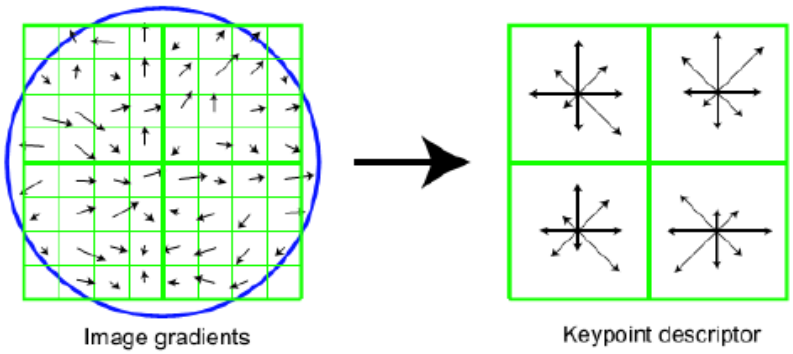



Figure 3.8: Left: the gradient magnitude and orientation at a sample point in a square region around the key-point location. These are weighted by a Gaussian window, indicated by the overlaid circle. Right: The image gradients are added to an orientation histogram. Each histogram includes 8 directions indicated by the arrows and is computed from 4x4 sub-regions. The length of each arrow corresponds to the sum of the gradient magnitudes near that direction within the region. [55]

### 3.4.5 Invariance Properties of SIFT

Scale Invariance is achieved for SIFT key-points by extracting and selecting the point of extremas which remain stable across the whole scale-space of DoG images. In order to achieve rotation invariance for the key-point, the coordinates of descriptor and gradient orientations are rotated relative to that of the key-point orientation. To make the key-point illumination invariant, the feature vector for the descriptor is modified. The illumination variations could be of several types and therefore the feature vector is modified accordingly. For example, a change in image contrast in which each pixel value is multiplied by a constant will multiply gradients by the same constant; so this contrast change will be cancelled by vector normalization to unit length. A brightness change in which a constant is added to

each image pixel will not affect the gradient values, as they are computed from pixel differences. Therefore, the descriptor is invariant to affine changes in illumination. However, non-linear illumination changes can also occur due to camera saturation or due to illumination changes that affect 3D surfaces with differing orientations by different amounts. These effects can cause a large change in relative magnitudes for some gradients, but are less likely to affect the gradient orientations. Therefore, influence of large gradient magnitudes is reduced by thresholding the values in the unit feature vector, and then renormalizing to unit length.

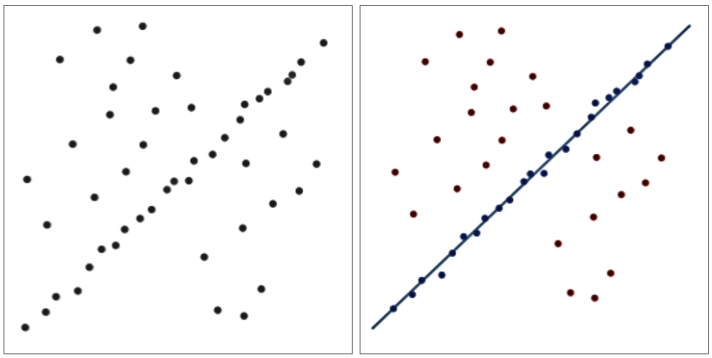
## 3.5 RANdom SAmple Consensus (RANSAC)

**RANSAC** is a method to approximate factors of a calculated model from a data set which includes outliers. It can generate practical results only with specific probability, with this probability increasing as more iterations are allowed. This algorithm was proposed in [60].

It is assumed that the input data set consists of inliers and outliers. Distribution of inliers can be described by a model where the outliers do not fit the model. The outliers can be caused by noise or incorrect measurements. RANSAC assumes that there is a formula which can approximate the inliers, from a model that clarifies or fits this data.

Consider a dataset of points which contains inliers and outliers (Figure 3.9). When attempting to fit a two dimensional line to this dataset, a minimum square method

can be applied on the points for line fitting. Outliers will be considered as of having an inappropriate fit as they will not fit to the line. RANSAC creates a model using the inliers only, as long as the probability of picking inliers in the dataset is sufficiently high. It is not guaranteed that this condition will be satisfied, yet there are several parameters in the algorithm which needs to be selected carefully to retain the probability level realistically high.



(a) (b)

Figure 3.9: (a) dataset of points including outliers for which a line will be fitted. (b) Fitted line after application of RANSAC; it is clearly visible that outliers do not have effect on the result. [82]

## 3.6 Local Energy Shape Histogram (LESH)

Feature Extraction is a very important stage in any VMMR system. Features need to be distinctive enough to distinguish between a large number of vehicle makes and models; at the same time, they also need to be invariant to changes in scale, colour, and illumination. These features should be able to model the shape of the local components (grill, headlamps) in the region of interest. The proposed system makes use of Local Energy Shape Histogram (LESH) first introduced by [65], which was

initially suggested at the start of the project and performed satisfactorily when compared with other feature descriptors. The LESH feature vector is extracted from the region of interest, and normalized between [0, 1] (Figure 3.10). The theoretical aspect of LESH can be explained as follows.

The local energy model was first proposed by [66]. It postulated that features are perceived at points in an image where the local frequency components have maximum phase congruency. The Local Energy indicates the corners, contours, or edges of underlying shape in an image [66]. The local frequency information is extracted by convolving a bank of Gabor wavelet kernels created using five spatial frequencies and eight orientations with the image; thereby retaining the underlying phase information. This is shown in equation below:

$$R\left(e_{s,t}, o_{s,t}\right) = I(x, y) \times \psi_{s,t}(x, y) \qquad 3.13$$

where $s, t$ is the scale and orientation respectively, $\psi$ is the bank of Gabor wavelet kernel, $I$ is the image with position $(x, y)$ and R is the response at image position $(x, y)$, the response consist of a real and imaginary part. The amplitude $A_n$ and phase $\phi_n$ of the response is computed using the equation below.

$$\begin{cases} A_n = \sqrt{e_t^2 + o_t^2} \\ \phi_n = tan^{-1}\dfrac{e_t}{o_t} \end{cases} \qquad 3.14$$

LESH makes use of the extended local energy model given in Equation 3.15, where is a constant used to avoid division by zero, $T$ is the estimated noise influence, $W$ is

the weighting of the frequency range. A more detailed explanation of this model is available in [66].

$$E = \frac{\sum_n W(x) \left\lfloor A_n(x)(\cos\left(\phi_n(x) - \bar{\phi}(x)\right) - |\sin(\phi_n(x) - \bar{\phi}(x))| - T}{\sum_n A_n(x) + \varepsilon} \qquad 3.15$$

The image is partitioned into an *n* number of sub regions*;* then an orientation label map $L$ is created by assigning each pixel the label of the orientation at which the local energy is maximal at all scales. A local histogram of the energy is accumulated along each filter orientation for each sub region. The local histograms are extracted using the equation below.

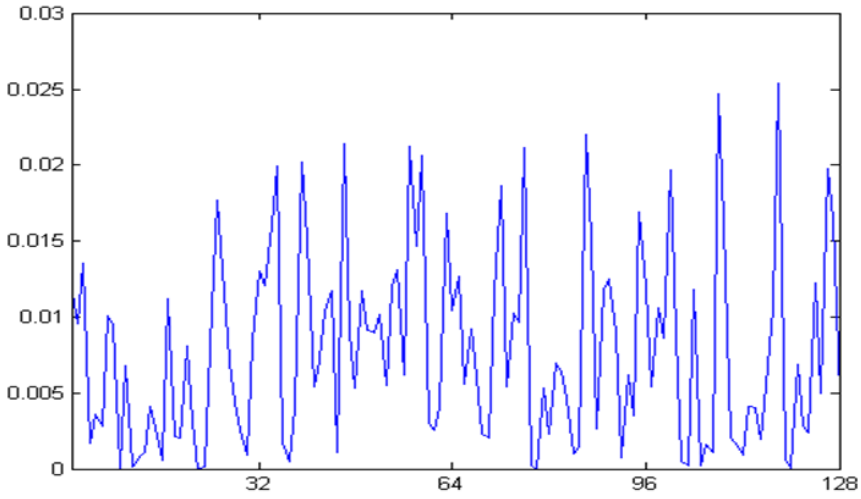
$$h_{r,b} = \sum w_r \times E \times \delta_{Lb} \qquad 3.16$$

Where $\delta_{Lb}$ is the Kronecker delta, $E$ is the local energy, $L$ is the orientation label map, $b$ is the current bin and $w_r$ is a Gaussian weighting function centred at region $r$ (Equation 3.17).

$$w_r = \frac{1}{2\pi\sigma} e^{\frac{\left[(x-r_{xo})^2 + (y-r_{yo})^2\right]}{\sigma^2}} \qquad 3.17$$

All local histograms are joined together to maintain a spatial relationship between local parts of the image. This will lead to an $n \times 8$ –dimensional feature vector (8 is the number of orientations). If the image is divided into 16 partitions, this will give a 128-dimensional feature vector.

(a) (b) (c)




(d)

Figure 3.10: Example of the LESH feature descriptor: (a) Original Image, (b) Local Energy Map, (c) Orientation Map, (d) LESH Feature Vector.

## 3.7 Data Association

The data association stage is important in tracking multiple objects. It enables the tracker to reconcile measurements from an object in a new frame at time $k$ to the measurements from the previous frame at time $k-1$ [83]. The data association algorithm is given below:

First the Euclidean distances from each of the centre point measurement(s) $n = (n_x, n_y)$ of the number plates in the current frame to every measurement(s) from

the previous frame $p = (p_x, p_y)$ is calculated. Equation 3.18 is the formula for the Euclidean distance.

$$d(p,n) = \sqrt{(n_x - p_x)^2 + (n_y - p_y)^2} \qquad 3.18$$

Then the distances calculated above are used to assign a single measurement from the current frame to its closest measurement from the previous frame. If the distance between a current measurement and its assigned previous measurement is greater than a predefined threshold the current measurement is assigned to a new estimator and the previous measurement is assumed to have been undetected either due to occlusion or leaving the frame. If no previous measurements exist, the current measurement(s) are considered to be new vehicles. New estimators are created for these measurements.

# CHAPTER 4

# Speaker Identification for Video Conferencing

## 4.1 Introduction

One of the significant practical problems in video conferencing facilities widely available at present are the difficulty in determining who is speaking amongst a large number of would be conference participants. There is a need for providing users (i.e. conference participants) automatically with a close-up of the current speaker which continues to automatically track a new speaker when one begins to speak. A number of attempts have been made in literature to address the above issue.

In [9] a speaker identification system was presented that was supported by dynamic visual data from video sequences including the lip region. The geometrical features of the lips were analysed included data about the shape and intensity information of said lips. This person identification system was based on a Hidden Markov Model (HMM). In [1] authors used morphological filtering for localization and an adaptive template matching for facial feature extraction for identity confirmation and in [13] authors used visual information from different components in the face for speaker

identification. The research of [14] presented a system using multimodal visual information from a video sequence. The modalities, face, voice and lip movement were merged using voting and opinion synthesis. In [15] a dynamic video-only biometric system was implemented with a robust and automatic method for tracking a speakers' face.

Despite all the above efforts speaker identification still remains an open research problem due to issues related to speed and accuracy of the existing systems. Real-time techniques capable of performing well under different levels of facial illumination, shadows etc., especially on the lip region which is very closely analysed for possible detection of movement, is required. Our approach summarized below attempts to extend the state-of-art in speaker detection by proposing a robust, real-time system.

## 4.2 An Overview of the Proposed System

In a typical video conferencing environment, multiple microphones and cameras located in multiple meeting rooms capture audio signals and video streams. The audio signals arriving from the microphones can be easily analysed and classified as human voice or non-voice audio (e.g. noise, music etc.) using a human voice detection approach. The presence of the human voice in a channel verifies that the associated video stream would contain the speaker. However a video channel (i.e a given node in the videoconferencing system) can contain many individuals out of whom only a particular person is talking. Therefore the selected video stream can be

first analysed with a facial detection algorithm to determine the presence of human figures. This algorithm can locate faces in each frame of the video stream using Haar like features. The face detection stage can subsequently be followed by a lip localization algorithm. It may be assumed based on human facial anthropometric data that the human mouth is located in the bottom third of a human face. Therefore a simple subdivision of the facial area can lead to lip localization. The last step is to detect the lip movements, which is a challenging task. The basic idea of the lip movement detection algorithm is to compare corresponding points between two views of a lip and measure the relative movements and subsequently classify as to whether the shape of the lip has changed significantly, i.e. whether the face indicates an individual who is talking. In the proposed work Coherent Point Drift (CPD) algorithm is used (see Chapter-3) for the above purpose. For each comparison, variance is calculated between the locations of the corresponding points. As the average variance for ‘not talking’ frames is considerably lower comparing to the ‘talking frames’ a robust classification can be obtained. Due to the robustness of the CPD algorithm to noise and limitation of edge detection algorithms, It is proven that the proposed approach to lip movement detection is robust, whilst performing real-time.

Section 4.3 presents the proposed system; sections 4.3.1, 4.3.2, 4.3.3 include detailed discussions on face detection, lip localization and lip movement detection approaches subsequently utilized in proposed design/implementation. Section 4.4 provides experimental results and an analysis. Section 4.5 concludes with suggestions for further improvement.

## 4.3 The Proposed System

Figure 4.1 illustrates the block diagram of the proposed speaker detection system, which includes the audio processing blocks that are not discussed in detail in this research work as they are assumed to be outside the scope of research presented in this thesis. The speaker identification system starts by analysing audio signals which come from different channels. The signals are checked and categorized as human voice and noise using a pre-trained algorithm that is capable of identifying human voice from other audio. As a result of this audio processing the video stream through which the human voice comes from is selected. The selected video is analysed for speaker identification (amongst the possible presence of many human faces) using three key stages of computer vision algorithms as described in Figure 4.1.

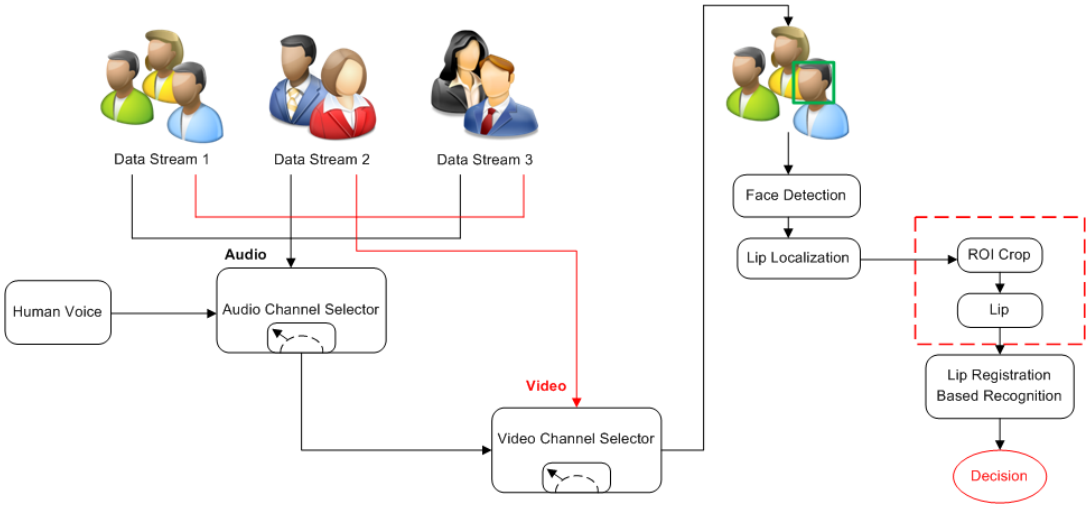



Figure 4.1: The proposed speaker identification system

### 4.3.1 Face detection

Face detection is the first step in automatic face recognition. In the proposed speaker identification system for this purpose the well established and widely used OpenCV face detector was used [70]. In this approach to face detection the aim is to build a cascade classifier with high correct detection and low false detection rate, although every stage in the cascade has a strong classifier including a group of weak classifiers. All weak classifiers are bounded to a unique Haar-like feature. To select the specific Haar features to use, and to set threshold levels, [64] have used AdaBoost [61]. It is noted that the AdaBoost procedure is used to choose more appropriate features to form a stronger classifier. It merges a number of weak-classifiers to create one strong classifier. A weak classifier is the classifier which gets the correct answer more often than random guessing. Each of the weak classifiers adds more correction to final answer. A classifier is trained with a number of different views of a face, which are balanced to a unique size and positive and negative samples - arbitrary images of the same size. Two sets of grayscale sample images were used in the proposed system's training procedure. One set included human face samples and the other included random non-face samples.

After training, the classifier can be applied to a region of interest in an input video or image. The classifier output will return a value of "1" if the detected region contains a human face; otherwise it will return "0". The search is based on the "window search", which moves across the whole image or each frame of video and

checks each and every window location using the trained classifier. Viola and Jones named this filtering sequence a cascade classifier [63].

Our initial experiments revealed that the use of the OpenCV based face detection algorithm described above could not satisfy the expected levels of accuracy, when used directly on some input video streams due to differences in colour representation formats. Some faces were not detected and some others were falsely detected confirming the need to include a pre-processing stage to convert each input video stream to a standard video colour representation format, i.e., 4:2:0 YCbCr (YV12) [75].

Despite the above attempt, during the execution of the face detection process, it was revealed that some non-face areas were falsely detected as face areas. However a close analysis of such false detections revealed two important facts. Such detections are not continuous, i.e. detected in a significant number of consequent frames, but are rather spontaneous. Further as the focus of the capture system (manually operated or automatic) is human faces, it is possible to assume that detected human face-like regions should be significantly large. Therefore false detections can be further reduced by removing small, spontaneously detected regions detected as faces, from further consideration. Therefore in the proposed system a further image processing and filtering stage was included to resolve the issue discussed above. In each frame two points are defined for identifying each rectangle that encloses a region that is likely to contain a face following the face detection approach described above. They are the top left and the bottom right corners of the rectangles. Using the co-ordinate values of the above points the areas of the rectangles are calculated and the values which generated rectangles above a defined size threshold

are picked up for further consideration (i.e., small regions are ignored). Subsequently assuming that a human face should appear consistently within reasonable displacements between consecutive frames, the movement of each rectangle over the frames of the video sequence is monitored. When system starts it makes an evaluation about the distance of the participants from the camera. If the participants are too close or too far from the camera a cropping calibration will be needed to zoom in or zoom out on the participant to bring the faces to the desired distance. After calibration in a video which frame rate is 30 per second, it is assumed that the location of the face will not be moving more than about 20 pixels (in any direction) due to practical reasons, in consecutive frames. Thus the rectangular regions in current frame will be in a neighbouring location to the corresponding rectangular region of the previous frame. Thus the position of each rectangle's two reference points (i.e., x and y values of top left and bottom right points of rectangle) in the current frame is compared with those points of the rectangles in the previous frame. If the difference is bigger than 20 pixels and this remains to be the case for 5 consecutive frames, the corresponding rectangles are removed from being considered further. The overall idea is to remove outliers of the facial locations within a video based on the assumption that the facial movements in a conversational video (i.e. video conferencing) are smooth.

### 4.3.2 Lip localization

Next step is lip localization. Lip movement detection initialization involves an approximate estimation of the lip location. The lip locator has to initially find the location of the mouth in a human face in a digital image or a video frame.

Most lip localization techniques are feature or edge based. Disadvantage of feature based approaches is their low speed. Numerous features must be tested on several facial parts before a correct location of the mouth can be found. Thus these techniques will not be suitable for implementing real time systems when added to the complexities of the other important stages of a speaker detection system. Lips can also be found using edge detection. The reason of detecting sharp changes in image brightness is to notice important changes in properties of the image. The outcome of applying an edge detector to an image will be a set of linked lines that specify the borders of objects. Hence, applying an edge detector to an image considerably decreases the quantity of information to be processed and as a result eliminates the data that in less important, while keeping the significant information of the image. Unfortunately, edges detected in images can often lead to poor results due to the edge lines being not connected, i.e. fragmented, and also due to false detections or missed edges. Thus edge detection based approaches for lip localization is not robust.

Due to the above reasons for lip localization a simple approach based on human facial anthropometric ratios is adopted. Once the face detector detects the facial region, it is subdivided horizontally into three regions and the lips are assumed to be

located in the bottom third. In a similar way, the face is divided by four vertical lines the lips are located horizontally to be within the two middle quarters of the face (see Figure 4.2). Based on our experiments this approach was both simple enough to be real-time in operation and robust to different conditions like poor lighting, etc. Once the lip area is localized more detailed analysis of the smaller space can be used to detect possible lip movements. This will not be computationally extensive as the area being processed is substantially smaller as compared to the facial area.

Using the above approach, the lip location can be extracted from the detected face. As a result the lip area will be localized and can be used in the next step, i.e., lip movement detection.

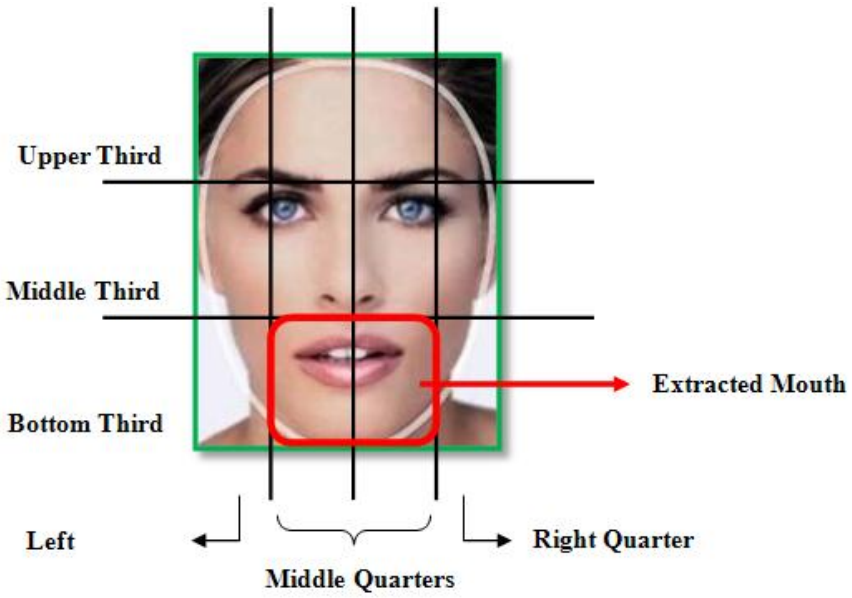
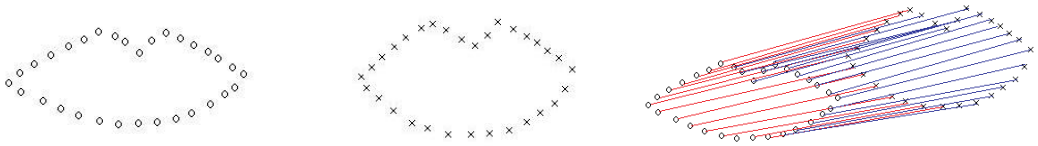



Figure 4.2: Facial shape and anthropometric division

### 4.3.3 Lip Movement Detection

This stage enables the identification of a talking person by comparing of the shape of the lips to a closed mouth or via temporal analysis of shape changes that occur in a lip within a sequence of frames. The approach proposed in this thesis is based on the idea of CPD [54] (see Chapter-3).

The CPD idea is utilized in the proposed work for lip motion detection. The idea is to first carry out an edge detection (e.g. Canny edge detector) on the localized lips of all frames and to compare the edge points of all lips in a set of frames with the edge point set of a reference non talking lip and check the amount of change between their correspondences based on the CPD approach (see Figure 4.3). If the variance in the deviations is low, it means that the lip has not moved or the movement cannot be considered as sufficiently large to classify the lip as a talking lip and if the result is high, it means that the lip shape has been changed signifying a talking lip. For achieving this idea, the first frame of video (which is assumed to be a non-talking lip) is chose as the reference image. Then using the canny edge detection the edges of the lip will be taken out and saved as the point sets. For the images which will be registered to the reference image the same action will be performed to obtain the point sets. All upcoming frames will be compared to the base image.



(a) (b) (c)

Figure 4.3: (a) Point sets of non-talking mouth, (b) Point sets of talking mouth, (c) Correspondence distance between point sets

Each point from the second image will be moved to the base point and this movement will result in a set of correspondence distances. Then, the variance of correspondence distance is calculated for all frames. Plotting and comparing the variance of whole video, it can be easily seen that there is a significant difference between the results of talking lips and non-talking lips. Obviously if the person is not talking the lips will not move significantly. Thus the correspondence distance will be small and vice versa. By calculating the Mean of all variance which resulted from annotated training videos it was found that average a variance values below 15 signifies a non-talking lip whereas above this threshold it can be classified as a talking lip. The threshold 15 was determined empirically and can be different in other experimental data sets.

## 4.4 Experimental Results and Analysis

A total of ten video samples consisting of 750 frames were used to evaluate the performance of the proposed approach to speaker identification. From the selected test videos, every fifth frame is used in order to reduce CPU usage and manage

memory utilisation problems. Experiments showed that in a 30 frames per second video, there will not be significant changes between the video frames. These samples depicted conversational videos obtained under various illumination conditions, included people with different skin colour/tone, still/moving faces, talking faces / non-talking faces or a combination of the above mentioned situations.

### 4.4.1 Analysis of face detection

Different videos as described above were used to test the OpenCV face detector utilised as the first stage of the proposed approach to speaker identification. As mentioned previously (see section 4.4) this approach resulted in some faulty detection (see Figure 4.4).

(a) (b) (c)

Figure 4.4: (a) Faulty face detection (i.e. the outer rectangle) using the OpenCV face detector [70], (b) Faulty face detected using the false detection identification approach proposed, (c) Revised result for face detection.

It can be seen from the sample result illustrated in Figure 4.4 more than one rectangle was identified as to include a face when the original face detection approach was used. The first step was to eliminate false detections and leave only rectangles that enclose possible faces. For this reason, the extended algorithm proposed in section 4.4 that eliminated false face detections was utilised. As a result all false detections are eliminated and the correct faces are included in a rectangle marked green, rather than the markings indicated in red in the first pass using the original face detector.

### 4.4.2 Lip localization

In order to localize the lips, the detected face was cropped by separating out the bottom third of face (horizontally) and middle two quarters (vertically) (see section 4.3.2). Figure 4.5 indicates the extracted lip areas from the detected face.

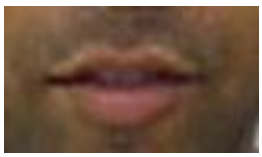


Figure 4.5: Successfully cropped lip area

### 4.4.3 Lip movement detection

The results from lip localization were used in the lip movement detection stage. The first frame which was assumed to be a non-talking lip was chosen as the base image. Then using the Canny Edge Detector, the boundaries of the lips were found

(see Figure 4.6). The same process was applied to all frames. Note that the edges are represented by a set of pixels; therefore, they can be used as point sets which can subsequently be analysed using the CPD based approach.

(a) (b)

Figure 4.6: Lip detection (a) Sample of cropped lip area, (b) Detected lip boundaries

The challenging task in this section was edge detection. The Canny Edge Detector did not perform as expected in some images. Setting the threshold was difficult as finding the best threshold (i.e., which gives the best detection of edges) to be used accurately in each frame would have resulted in significant overhead. Therefore a single threshold range was used for all the frames which at times resulted in some non-optimal edge detections such as broken edge representations and some edges not been picked up etc. However the CPD approach automatically detects the sets of corresponding pairs of points between the edge representation of the test and reference frames, ignoring edge points which do not find a corresponding edge point being detected in the other image. The edge point set of each detected lip in frames was compared to the edge point set of the reference image's lip and the correspondence distance between two point set was calculated. For each lip a set of correspondence distances are obtained. The variance of each point set was calculated and plotted. Observing this plot (see figure 4.7, 4.8 and 4.9) talking and no-talking lips can be identified by checking against an average variance threshold

value of 15. Lips that resulted in average variances above 15 were classified as talking lips whereas below were considered as no talking lips.

The following tests were done on video samples that included as the only person, a talking and none talking individual. The table 4.1 tabulates the variance of position of each rectangle within the relevant frame of the video consisting of the none-talking individual. The graph in figure 4.7 illustrates the variations clearly. It can be easily seen that the variance of each frame for the non-talking video remains low. It is noticeable that the mean of variances for a non-talking video is always below 15. The Table 4.2 tabulates the variance of location of the rectangles within the relevant frames of the video containing the talking individual. The corresponding graph illustrated in Figure 4.8 draws attention to the fact that high values (above 15) of average variances are depicted in the case of the video containing the talking individual.

**Variance of location of rectangle for the video containing the non-talking individual**

| | | | | |
|---|---|---|---|---|
| 0.4369 | 9.5727 | 7.9871 | 1.0750 | 3.7762 |
| 4.9819 | 6.6978 | 11.9011 | 6.0395 | 4.6752 |
| 7.8971 | 5.5245 | 9.0711 | 9.5866 | 8.7358 |
| 5.0170 | 10.7884 | 1.5821 | 8.0947 | 6.6613 |
| 7.9375 | 10.9437 | 8.3186 | 4.9978 | 4.4416 |
| 5.9088 | 8.7256 | 9.6574 | 7.5840 | 1.2785 |
| 7.1988 | 11.4776 | 7.1016 | 6.9594 | 5.1536 |
| 8.6245 | 6.3971 | 8.0722 | 10.0020 | 8.9503 |
| 6.4699 | 3.2315 | 8.8776 | 11.0709 | 7.0293 |
| 4.3187 | 1.3426 | 7.6989 | 9.5097 | 4.7386 |
| 5.1851 | 4.2939 | 1.5729 | 8.7676 | 4.0468 |
| 5.0251 | 0.4130 | 1.0393 | 7.8786 | 5.0150 |
| 8.9503 | 1.6902 | 0.6533 | 6.3959 | 4.1491 |
| 7.2586 | 3.6085 | 3.8014 | 6.5985 | |
| 7.5284 | 10.3956 | 1.3067 | 9.3396 | |

**Mean of Variances : 6.2881**

Table 4.1: Variances of location of rectangle – video containing the non-talking individual

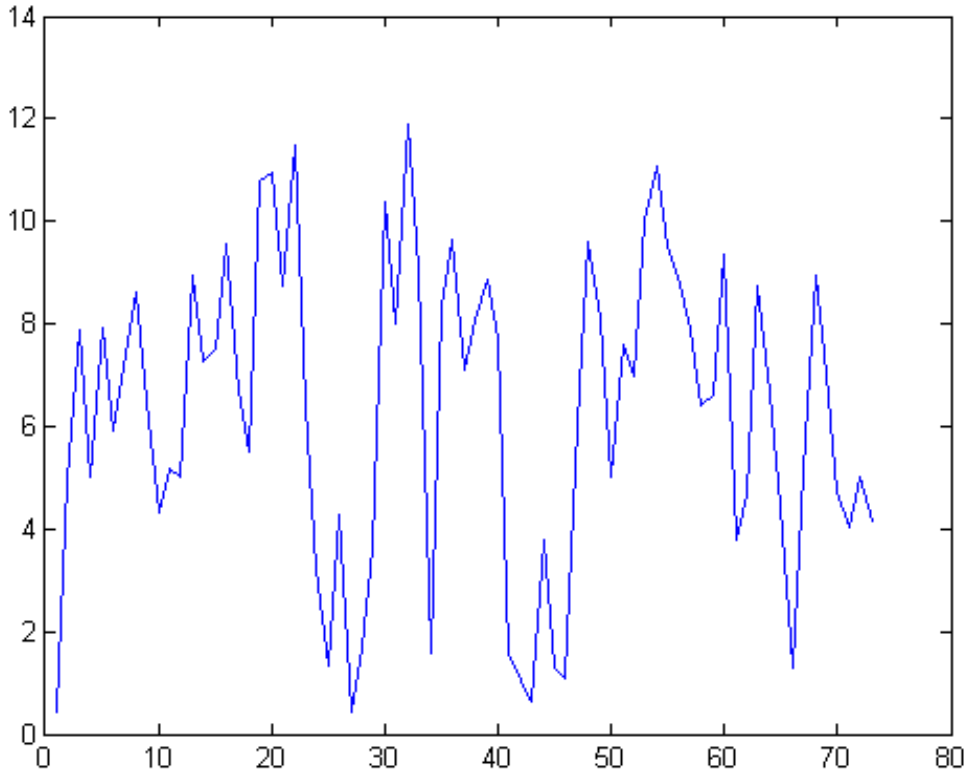



Figure 4.7: The variance plot for the video consisting of non-talking individual

**Variance of location of rectangle for the video containing the talking individual**

| | | | | |
|---|---|---|---|---|
| 26.209 | 28.5329 | 37.5262 | 34.8882 | 39.8508 |
| 34.1631 | 21.1533 | 32.1238 | 36.4393 | 33.9797 |
| 19.6882 | 23.5028 | 26.8475 | 30.6281 | 24.0042 |
| 27.7730 | 22.5713 | 22.6922 | 21.7208 | 19.7109 |
| 21.3237 | 19.3798 | 20.8355 | 18.9148 | 18.7812 |
| 18.8941 | 19.5306 | 20.1097 | 21.7513 | 19.8901 |
| 18.1954 | 27.5239 | 36.8231 | 23.3371 | 25.0352 |
| 10.7692 | 16.5397 | 19.2144 | 29.9694 | 38.7377 |
| 46.3544 | 40.7686 | 48.0673 | 43.5273 | 23.9449 |
| 31.4527 | 27.0735 | 36.6982 | 22.3877 | 16.4230 |
| 19.4384 | 19.8182 | 22.8714 | 11.1962 | 24.5385 |
| 13.2058 | 10.2361 | 25.7002 | 23.7931 | 23.5954 |
| 31.7812 | 30.3110 | 27.5770 | 38.3885 | 47.6030 |
| 25.1865 | 26.5777 | 23.7654 | 33.7151 | 39.6288 |
| 31.0835 | 20.6688 | 11.0061 | 25.9198 | 22.9828 |
| 11.3495 | 10.6443 | 15.8840 | | |

**Mean of Variances: 25.8287**

Table 4.2: Variances of location of rectangle – video containing the talking individual

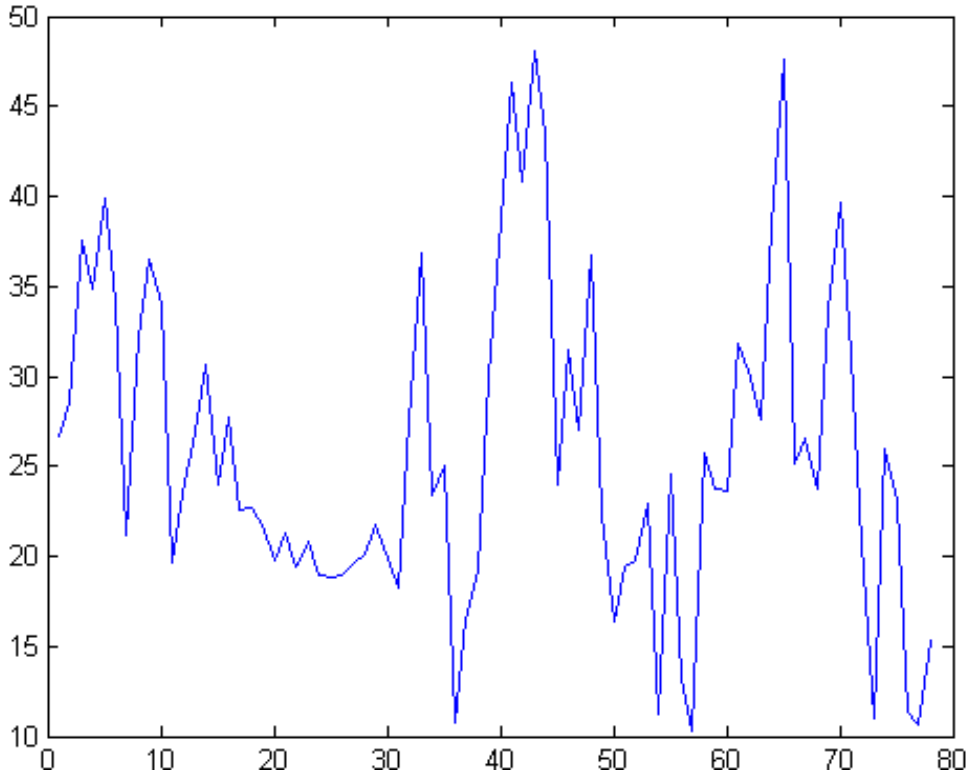



Figure 4.8: The variance plot for the video consisting of the talking individual

The following results are from video which contains a person who first does not talk and then talks for a short while before stopping again. The divisions of talking and non-talking frames were as follows: fames 1 to 30, not-talking, frames 31 to 114 talking and frames 115 to 150 not talking. Looking at the values of variances from the Table 4.3 and graph in Figure 4.9 closely, it is seen that our thresholding approach allows the proper identification of the lip status change locations. Note that sudden unexpected changes in the variances are related to the incorrect edge detections or imperfectly cropped lip locations, which includes some points selected from the background. Note that the averaging of the variances before comparing with a threshold resolves this complication. It should be mentioned that for eliminating the false peaks, it is necessary to apply a low-pass filtering.

**Variance of location of rectangles for a video containing an individual who moves from being "Non-Talking, Talking to Non- Talking"**

| | | | | | |
|---|---|---|---|---|---|
| 12.5402 | 5.8915 | 11.2944 | 3.4690 | 6.5573 | 13.5782 |
| 4.0439 | 9.2163 | 3.1380 | 11.5830 | 4.8683 | 19.2469 |
| 2.3534 | 10.1023 | 5.9144 | 2.7443 | 4.3828 | 12.9343 |
| 3.1969 | 2.3832 | 7.4982 | 22.6378 | 3.4239 | 4.1534 |
| 10.7955 | 11.9025 | 9.3565 | 16.5893 | 11.2946 | 24.0416 |
| 40.3177 | 10.2719 | 38.1668 | 25.7490 | 27.6025 | 16.4222 |
| 19.4198 | 20.9311 | 23.5314 | 21.1017 | 18.0436 | 23.6091 |
| 19.7846 | 21.9171 | 19.6500 | 17.2726 | 27.7067 | 28.9641 |
| 26.1448 | 27.2893 | 28.9795 | 23.4561 | 34.9832 | 24.3939 |
| 19.2408 | 35.4537 | 33.8603 | 34.3512 | 25.6864 | 25.2074 |
| 29.7674 | 23.0222 | 19.9895 | 25.0900 | 23.9646 | 22.2552 |
| 26.0507 | 28.7443 | 27.8488 | 31.8092 | 23.3291 | 38.5097 |
| 37.4096 | 28.1230 | 23.5402 | 23.2991 | 39.2026 | 34.8993 |
| 41.9703 | 33.7636 | 29.7331 | 28.9408 | 32.7239 | 47.6175 |
| 24.4182 | 19.2816 | 26.4819 | 25.7852 | 30.8299 | 19.8362 |
| 26.1384 | 31.5121 | 37.7677 | 27.5213 | 17.4678 | 33.2208 |
| 23.9804 | 25.0853 | 35.4589 | 24.5779 | 33.7198 | 27.3530 |
| 21.2638 | 44.9583 | 24.5016 | 23.6719 | 31.4601 | 19.0088 |
| 21.6824 | 19.4404 | 18.9456 | 17.1437 | 35.6352 | 16.0351 |
| 13.5663 | 28.9866 | 17.5990 | 22.6257 | 9.9878 | 8.2797 |
| 5.9969 | 11.4613 | 5.9726 | 4.5866 | 10.3414 | 4.9446 |
| 5.0051 | 4.6294 | 3.5961 | 3.1720 | 2.1722 | 3.9540 |

| | | | | | |
|---|---|---|---|---|---|
| 10.5736 | 8.1813 | 15.1029 | 2.7212 | 9.3454 | 13.6903 |
| 8.6685 | 3.1751 | 8.4914 | 4.2121 | 9.3949 | 11.4330 |
| 11.8122 | 11.0833 | 10.2734 | 11.3397 | 8.3568 | 8.8642 |

**Frames 1-30: Not Talking**

**Frames 31-114: Talking**

**Frames 115-150: Not Talking**

Table 4.3: The variances plot for a video containing an individual who moves from being "not-talking, talking to non- talking"

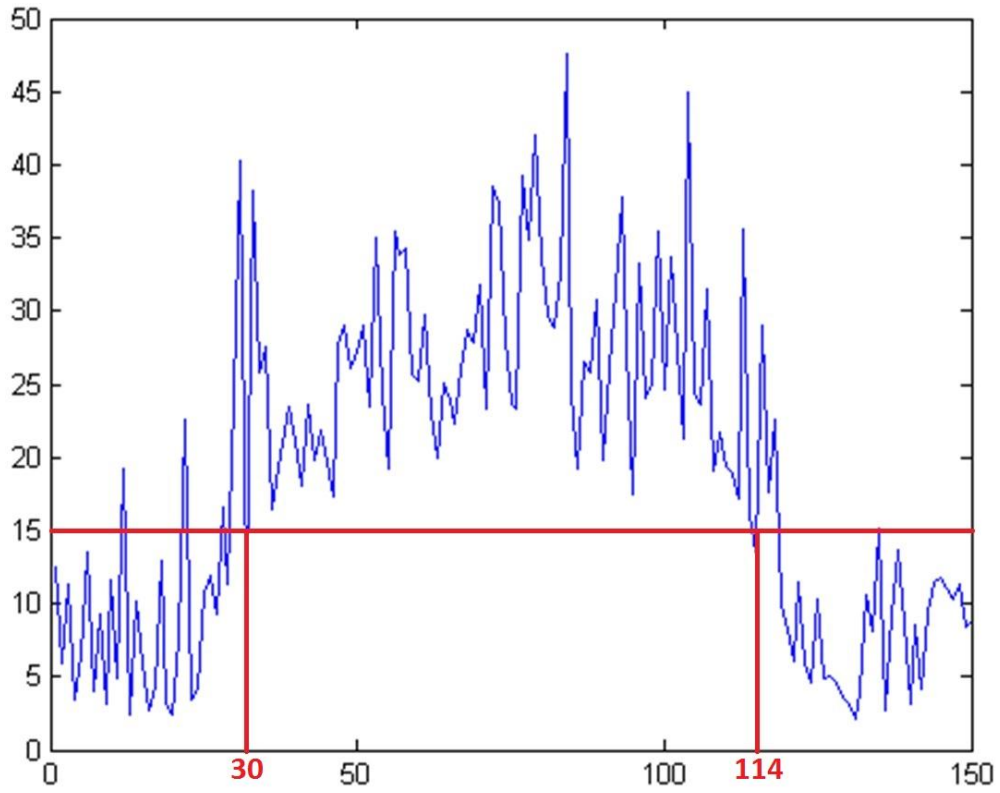



Figure 4.9: The variance plot for the video consisting of the talking and non-talking individuals. Some false peaks can be seen in the first 30 and last 30 frames, which can be eliminated by applying a low-pass filter.

Experiments were carried out on a set of 10 video sequences. The results were separated by manually separating the sequences to continues, known talking or non-talking clips, and also by separating out video clips that varied through a non-talking, talking and non-talking sequence of status. It was observed that for the former set of video clips the proposed approach achieved 97.2% rate of success and for the later set rate of success was 96.1%.

## 4.5 Conclusion

In this chapter a novel approach for speaker identification has been described, based on Coherent Point Drift (CPD). Three steps were carried out to achieve the goal; namely, face detection, lip localization and lip movement detection. An improved version of a commonly used face detection algorithm has been adopted in order to obtain the correctly detected faces per frame. A human faces typical anthropometric division / ratio was used in lip localization. A novel method was proposed for lip movement detection based on CPD which was used to compare edge point sets of a given face with the edge point set of a known, non-talking, reference lip. CPD concurrently finds the transformation and the correspondence distance between two different point sets. The CPD method is most effective when approximating still non-rigid transformations and illustrates robust and precise performance regarding noise, outliers and missing points. Experiments were carried out to test the proposed system using several video samples; the system attained 97.2% accuracy for the videos which were taken separately for non-talking and talking faces, and 96.1% for

the videos which were taken continuously varying through non-talking, talking and non-talking states.

As further work this approach can be integrated into a practical, widely used, video conferencing system (e.g. The UK access grid) to give the users the benefits intended by this novel approach.

# CHAPTER 5

# Multi-Exposure and Multi-Focus Image Fusion with Compensation for Camera Shake

## 5.1 Introduction

In the field of High Dynamic Range (HDR) imaging technology, in the past decade there have been significant developments as a result of increased consumer demand for experiencing images that are perceptually closer in appearance to images perceived by the human psycho-visual system. To this effect, HDR imaging sensors are at present replacing the traditional Standard Dynamic Range (SDR) sensors in digital cameras. However, the lack of developments in image/video encoding algorithms and display technology capable of making practical use of HDR images makes it still important to find alternatives to rendering HDR scenes using SDR imagery. Therefore, a number of algorithms have been proposed in literature to fuse multiple-exposure SDR images that result in images that are perceptually similar to HDR images, i.e. images perceived by the human eye. Multi-exposure image fusion involves the fusion of multiple consecutive images of the same scene taken at quick succession by a SDR camera. However, a problem thus arises: camera shake can cause severe de-registration between multiple images that invalidate the direct

applicability of many existing multi-exposure image fusion algorithms. The camera shake can be translational (vertical, horizontal) or rotational (in-plane and out-of-plane) and it is thus important that compensation for both types of shake is carried out prior to image fusion.

Similarly, due to the limited depth of field in optical lenses, it is usually impossible to capture an image that contains all relevant objects in focus. A solution for this is a multi-focus image fusion that fuses two or more images that are captured using different camera settings (i.e. different focuses) of the same scene in order to form a final image with uniform focus and sharp content. Due to the slightly different focus settings used in capturing the constituent images for fusion, in addition to the presence of camera shake in translational (vertical, horizontal) and rotational forms, compensation for the effect of image zooming, importantly requires to be considered.

The focus of the proposed research is the development of an end-to-end multi-exposure and multi-focus image fusion system that addresses the issues of camera shake, low dynamic range in SDR cameras, and the limited depth of field of optical lenses. A significant number of multi-exposure image fusion algorithms have been proposed in literature [20, 21, and 25]. However, only very few algorithms focus on the problem of camera shake [23, 35-38]. Furthermore they are severely limited in their ability to compensate for different types of camera shake.

Image fusion is the combination of images though a specific fusion algorithm, so that the resulting image is clearer and more intelligible. Image fusion can take place on pixel-level, feature-level, and decision level. In the literature, image fusion has been based on pyramidal fusion, Contourlet fusion and wavelet fusion. The

pyramidal approach could be considered as computational inefficient due to the redundancy presented in the pyramidal transform. The wavelet transform results are acceptable in natural images, but smooth edges cannot be detected powerfully because of its restricted three directions (horizontal, vertical and diagonal) to detect features in the images. Contourlet transform is a two-dimensional transform that has the capability to effectively represent images containing curves and features [57, 58]. In the Contourlet transform, multi-scale and multi-direction analyses are done separately. First, the Laplacian Pyramid (LP) transform is used to perform a multi-scale decomposition and then a Directional Filter Bank (DFB) is used to filter the high frequency components from each LP channel. Therefore a Contourlet-based image fusion method produces improved outcomes as opposed to wavelet-based fusion [24]. However the Contourlet transform's redundancy ratio is less than 4/3 because of the LP transform and due to the fact that the multi-resolution structures are not constant as the number of directions in DFB, which are variable. An approach proposed in [31] provides a solution to the above shortcoming, i.e., wavelet-based Contourlet transform (WBCT), which is non-redundant and has a multi-resolution structure. The advantages of using wavelet-based Contourlet transform are that it solves the problems of multi-scale localization, directionality and anisotropy. However, the WBCT fusion approach adds artefacts, and therefore the perception of the image fused, i.e. the details and intensity are highly affected.

Multi-focus image fusion algorithms have also been proposed [24, 33 and 34] and can be classified into two categories. (I) Pixel based methods: wavelet and Contourlet transform, spatial frequency and morphological operators, and (II) region based methods: multi resolution methods pyramid based approaches and

discrete wavelet based methods. The advantage pixel level image fusion is that images contain original data and therefore the pixel information is preserved. On the other hand, region based methods are very useful for image fusion because the real world objects usually consist of structures at different scales and human visual system [34].

In this research work, a new image fusion algorithm that allows compensating camera shake (translational, rotational and zoom) through a registration process, and a new approach of the WBCT fusion that allows combining aspects of both pixel level and region level (thus minimising artefacts) in order to fuse a set of images, are proposed. Section 5.2 provides a brief overview of the proposed system. Sections 5.3 and 5.4 present the operational and functional details of the proposed system, a detailed evaluation of the system's performance and experimental results of each step. Finally section 5.5 concludes with an insight to further work.

## 5.2 Proposed System

The proposed system consists of a registration module, which is used prior to image fusion. After the images have been registered and camera shake has been compensated, a fusion module is used in order to produce a single HDR image from a set of multi-exposure images, or a single multi-focus image from a set of images with different depths of field.

## 5.3 Image Registration

Figure 5.1 illustrates the block diagram of the proposed approach to registering a set of images captured in the presence of camera shake. The approach is based on the selection of a significant set of matching points, i.e. key points, between a selected base image and an image to be registered and subsequently using them to calculate the transformation matrix for image registration. Sections 5.3.1, 5.3.2 and 5.3.3 describe the three key stages of the image-registration approach adopted.

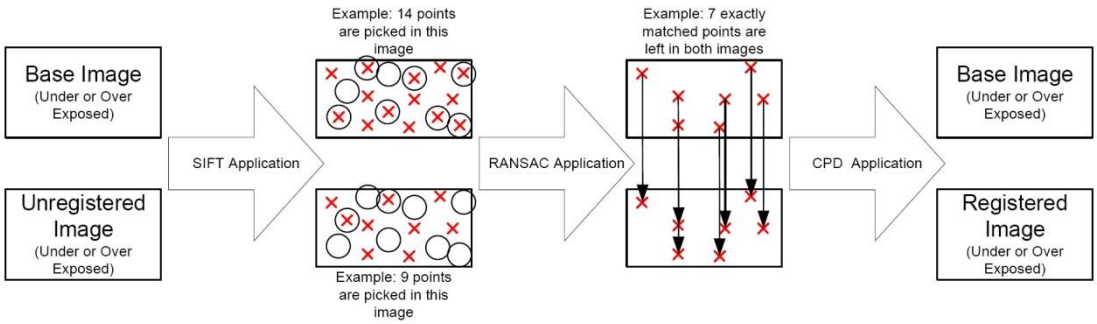



Figure 5.1: Image Registration Module

### 5.3.1 SIFT Based Key-Point Selection

The Scale Invariant Feature Transform (SIFT) [55] is an algorithm that is capable of detecting and describing local features of an image. It is invariant to rotation, scale and translation has made it a popular algorithm in many areas of computer vision and pattern recognition. SIFT is also partially invariant to illumination changes and robust to local geometric distortion.

In the proposed approach to image registration base image is selected from amongst the set of images in which another image from the set is registered to the base image. The algorithms first uses SIFT to find significant feature/key points in both images. In the case of multi exposure registration it is noted that the base image is considered to be the image with medium level of exposure amongst the multi-exposure image set when comparing the exposure changes that are represented within the image set being fused.

Due to the relatively large number of feature points that may be selected by SIFT; in carrying out the matching of key points between the base image and the image to be registered, it is likely that two geometrically non-corresponding points of the two images may match as they result in the minimum distance. Therefore reducing key point outliers prior to the matching key points will improve the reliability of matching and hence the outcome of the eventual image registration task. This requires the use of the stage that follows.

### 5.3.2 Using RANSAC to Remove Matching Point Outliers

The RANdom SAmple Consensus (RANSAC) [60] algorithm is an iterative method to approximate factors of a mathematical form, from a set of experimental data, which includes outliers. RANSAC is able to make robust estimations of the model parameters; it can estimate the parameters with a high degree of accuracy even when a significant amount of outliers are present in the data set.

The key point data sets generated from the SIFT stage consists of inliers and outliers. Within the purpose of using the RANSAC algorithm for the task at hand, outliers are defined as key points which are found to be present in the image being registered but not found in the base image in a specially, closer location. The outliers may result from poor illumination conditions, noise, etc. In proposed approach all SIFT key points resulting from the stage described in section 5.3.1 from the base image and the image being registered are first fed to the RANSAC algorithm. The base image matching points are assumed theoretically to be the inliers. Then RANSAC fits a model to these inlier points and tests the points from the image being registered against the fitted model, and if a point fits to the model it will be regarded as an inlier. The model is recalculated from each and every inlier, followed by an error estimation of the inliers relative to the model. The outlier key points are finally removed from the set of key points of the image being registered. The set of inlier key points of the image being registered are subsequently used for corresponding point matching with the set of key points of the base image.

As stated above, the removal of outliers using RANSAC results in increasing the reliability of subsequent SIFT key point matching. Thus the result of the stage described above is, two point sets consecutively from base image and the image to be registered which can now enter the final stage of a typical SIFT algorithm, i.e. key point matching. An Euclidean distance between feature vectors of key points is used to find corresponding point sets, by selecting all point sets whose distance is below a specified threshold which is determined experimentally.

### 5.3.3 CPD Algorithm for Registration

In this section Coherent Point Drift (CPD) [54] algorithm is used to register the images. CPD allows preparing the set of images to be fused in the subsequent stage of the proposed approach. CPD is based on 'Point Set Registration' and aims to form links between two given sets of points to find the corresponding features and the necessary transformation of these features that will allow the images to be registered.

In section 5.3.1 it was mentioned that the SIFT based approach to identifying key-points that can subsequently be used in matching allows the selection of key-points invariant to translation, rotation and scale. It was also mentioned that SIFT is also partially invariant to illumination changes and robust to local geometric distortion. Therefore the use of SIFT will allow the robust selection of potentially matchable key-points in the presence of camera shake (due to invariance to translation and rotation) and zooming (due to invariance to scale). Further the process will be partially invariant to illumination changes and effects due to the presence of geometric distortion. Given that in a typical HDR image capture and creation process of a handheld device/camera that involves fusing together a number of SDR images, captured under different exposure and focus settings, in quick time succession, may introduce relative vertical, horizontal, rotational movement, zooming, illumination changes and geometric distortion, the use of SIFT with abovementioned invariance are well justified. Further the use of non-rigid point set

CPD algorithm allows the proposed image registration approach to handle the presence of camera zoom between the constituent images.

Figure 5.2 illustrates images captured for the unit testing of the performance of the image registration module, which is applied prior to the multi exposure image fusion. The images were taken without the aid of a tripod, and the fusion was performed with two different image exposures in RGB domain. Figure 5.2(a) shows an overexposed image considered as the base image for the registration process and Figure 5.2(b) is the relevant underexposed image which is to be registered to the base image subsequently. The SIFT key feature points were found as a result of applying SIFT on the under-exposed and over-exposed images as illustrated in Figure 5.2(c) and 5.2(d). Subsequently using the RANSAC algorithm (Figure 5.2(e) and 5.2(f) the mismatched points are eliminated and finally by using the CPD algorithm the two images are registered. Figure 5.2(g) and 5.2(h) illustrates the key points before and after registration using CPD. To present a pre-view of the ultimate advantage provided by the abovementioned image-registration approach, Figure 5.2(i) and 5.2(j) illustrate, the output of the image fusion algorithm (see section 5.4) when the abovementioned image-registration approach is not adopted and adopted, respectively. In the case of fusing unregistered images (Figure 5.2(i)) it can be observed that the fused image appears blurry and smudgy in some parts of the image. Figure 5.2(j) illustrates the positive impact of prior image registration using the proposed algorithm. It can be seen that the subjective quality has being increased. Edges are sharper, no blocking artefacts were visible and the intensity or dynamic range of the final image has been optimised.


EXIT
FIRE EXIT
FIRE EXIT

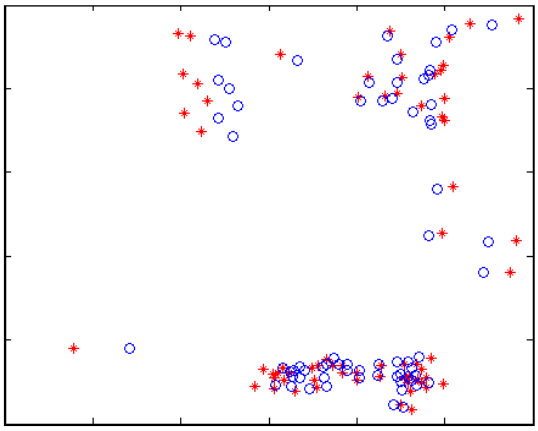
(a) (b)


FIRE EXIT
FIRE EXIT

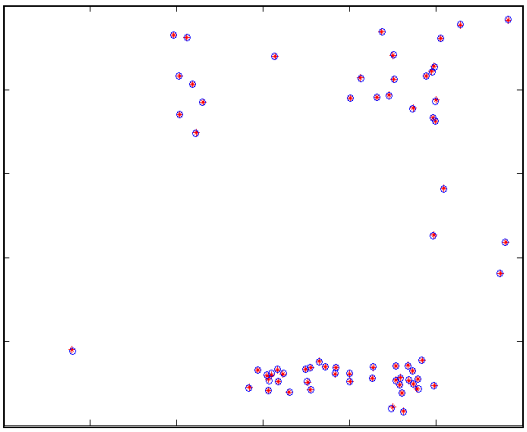
(c) (d)


EXIT
FIRE EXIT
FIRE EXIT

(e) (f)



(g)



(h)


(i)


(j)

Figure 5.2: Image registration process. (a), (b) Original set of multi-exposure images. (c), (d) SIFT selection of features in under exposed and over exposed images. (e), (f) Mismatched points eliminated with RANSAC in under exposed and over exposed images. (g) Key points from both exposures images (Blue circles from under-exposed and red crosses from over exposed images). (h) Key points matched. (i) Fused image without registration. (j) Enhanced fused image with registration

## 5.4 Multi-Exposure and Multi-Focus Image Fusion

Once all images are registered, a Wavelet based Contourlet Transform (WBCT) is used for identifying regions of maximum energy from within the multi-exposure or

multi-focus images to generate a fusion decision mask which is later used for fusing the multi-exposures or multi-focus images.

### 5.4.1 Wavelet Based Contourlet Decomposition

The Contourlet Transform [57] can represent the contours and textures in an image powerfully. The analysis of Contourlet transform typically takes place in two stages. Firstly, a Laplacian Pyramid is applied to achieve multi-scale image decomposition. Secondly, a directional Filter Bank is applied to obtain the high frequency directional sub-bands. However due to the use of the Laplacian Pyramid the Contourlet transform is redundant. In literature the use of Wavelet based Contourlet Transform, which is non-redundant has been proposed to resolve this problem [31].

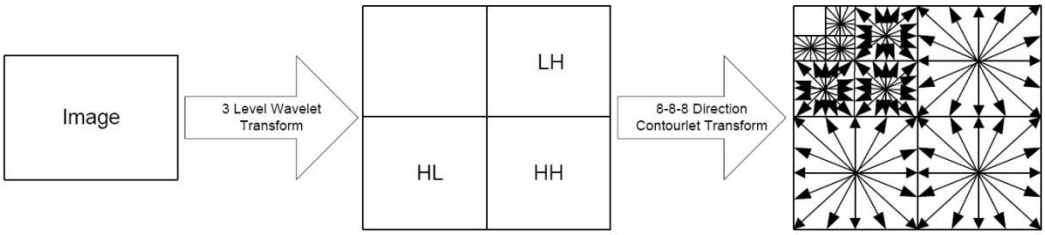



Figure 5.3: Frequency partitions obtained with WBCT

Figure 5.3 illustrates the frequency partitioning of a given image when WBCT is used. First, a three-level decomposition of wavelet transform is applied on an image giving nine high pass and one low pass sub-bands. Then a directional Contourlet transform is applied on each high pass sub-band of the wavelet decomposition. In experiments, 8 directions have been used to enable more detailed analysis. This

results in 8 directional Contourlet sub-bands and one low-frequency Contourlet sub-band for each high pass wavelet sub-band. Note that for clarity of Figures 5.3 and Figure 5.4 do not illustrate the low-frequency Contourlet transform. Figure 5.4 illustrates the block diagram of the fusion module.

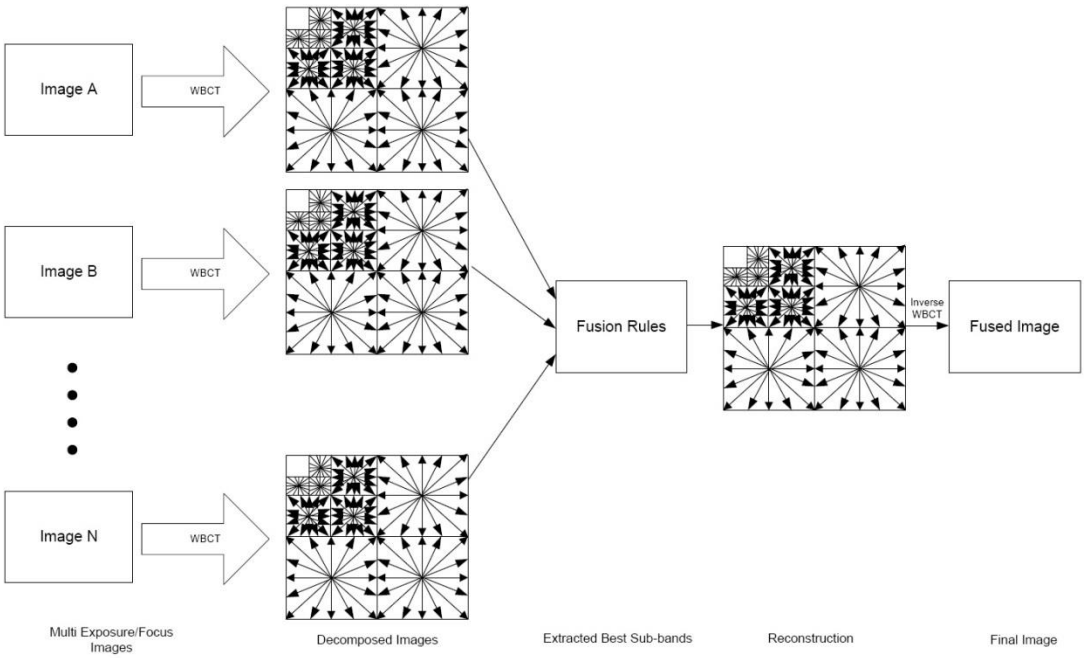



Figure 5.4: Fusion module

## 5.4.2 Improved Multi-Exposure/Multi-Focus Image Fusion

The basic idea of the fusion algorithm is to compare the corresponding sub-bands of the WBCT decomposition of each image of the multi-exposure image set and to determine the one with the highest energy, i.e. most detailed. Any attempt to fuse the sub-bands to obtain a fused image with best perceptual quality should consider the human psycho-visual system's sensitivity to different frequency ranges. The use

of different fusion rules is proposed depending on the frequency band of each sub-band being fused, as follows. In particular the use of two fusion approaches is tested.

### 5.4.2.1 Approach- 1

This approach uses three different algorithms to fuse together the Contourlet sub-bands (high and low frequencies) and the low frequency wavelet sub-band, formed when following the WBCT approach described in section 5.4.1.

***I - Fusion of High Frequency Contourlet Sub-bands:***

The high frequency sub-bands contain details of an image such as texture and edges whereas the low frequency sub-bands contains details of more spread-out nature or fuzzy in nature, such as background information. By calculating the absolute energy of high-frequency coefficients, the energy of a region can be obtained. A higher value means sharper changes. Region energy $E$ of a high frequency sub-band $E_H$ (where $H = (l, m, n)$, $l$-level of wavelet decomposition, $m$-LH, HL and HH bands of wavelet decomposition, $n$-directional Contourlet sub-band) of an image $I$ can be obtained as follows:

$$E_H^{(I)} = \sum_{(x,y)\in H} f_H^{(I)}\ (x,y)^2 \qquad 5.1$$

$f_H(x,y)$ is the coefficient at location $(x,y)$ of the high frequency sub-band $H = (l,m,n)$.

Considering that the sub-band of the multi-exposure image having highest detail will have the highest absolute energy, the H sub-band that contributes towards the fused image's H sub-band can be defined as:

$$f_H^{(F)}(x,y) = f_H^{(Y)}(x,y)\ ,\ E_H^{(Y)} = \max(E_H^{(i)})\ |_{i=1,2,\ldots n} \qquad 5.2$$

Where *n* is the number of exposures considered.

***II - Fusion of Low Frequency Contourlet Sub-bands:***

The low frequency sub-bands contain the fuzzy, spread-out information. Thus the fusion rule adopted is based on the region variances. The idea is to divide each low frequency sub-band to $i \times j$ rectangular sub regions and calculate the variance of each sub region which can be obtained as follows:

$$V_L^{(X)}(x,y) = \sum_{(x,y)\in(i\times j)} (f_L^{(X)}\ (i,j)(x,y) - \overline{f}_L^{(X)}(i,j))^2 \qquad 5.3$$

$\overline{f}_L^{(X)}(i,j)^2$ is the mean of all the coefficients in the rectangular sub region $i \times j$. Higher result in variance corresponds to more details. In the proposed approach-1 the fusion of the low frequency sub-bands are obtained from equation below:

$$f_L^{(F)}(i,j) = f_L^{(Y)}(i,j)\ ,\ E_L^{(Y)} = \max(E_L^{(i)})|_{i=1,2,\ldots n} \qquad 5.4$$

Where $n$ is the number of exposures considered.

***III - Fusion of Low Pass Wavelet Sub-bands:***

The low-pass wavelet sub-band of fused image is calculated as the average of the low pass wavelet sub-bands of the multiple exposure images, as follows:

$$A_L^{(F)}(i,j) = \frac{1}{n}\sum_{a=1}^{n} f_L^{(a)}(i,j)\ ,\ (i,j) \in LL \qquad 5.5$$

Where $n$ is the number of the multi exposure images and $f(i,j)$ is a coefficient from the low pass sub-band of the wavelet transform.

### 5.4.2.2 Reconstructing the Fused Image

After obtaining the best low and high frequency sub-bands of Contourlet transform and best low frequency sub-band of wavelet transform as above, the fused image is first formed in its transformed domain and later using inverse WBCT, is transformed to the pixel domain.

### 5.4.2.3 Experimental Results

Initially experiments were conducted on a standard set of multi-exposure images obtained with a stable camera set up (i.e. no impact due to camera shake) allowing the proposed approach to multi-exposure image fusion to be compared with many

existing multi-exposure image fusion algorithms. The results illustrated in Figure 5.5 prove that the proposed approach is capable of producing fused images of perceptually good quality.

Figure 5.5 illustrates images captured for the specific testing of the image registration stage, which is applied prior to the multi-exposure image fusion. The images were taken allowing the free movement of the camera, i.e. allowing shake. All images were Gamma corrected before processing by the proposed algorithms. Figure 5.5(a) is the base image (over exposed) and Figure 5.5(b) is an unregistered image of a different exposure setting (under exposed), which is intended to be aligned with the base image. Figure 5.5(c) illustrates the fusion result without prior registration of images indicating a blurry and smudgy nature on some parts of the image while Figure 5.5(d) illustrates the positive impact of prior image registration using the proposed algorithm. It can be seen that the quality has been increased in the form of increased sharpness and more details being observable.

A previous attempt to fuse images (Section 5.4.2.1) with the WBCT approach applied a three-level decomposition in a Laplacian pyramid (LP) [59] fashion using the wavelet transform, which allowed decomposing an image into nine high pass and one low pass sub-band [31]. Then, a directional Contourlet transform was applied on each high pass sub-band of the wavelet decomposition.

(a) Over Exposed (b) Under Exposed

(c) Fused without Registration (d) Fused with Registration

Figure 5.5: Experimental dataset

### 5.4.3 Improved Multi-Exposure/Multi-Focus Image Fusion

The WBCT decomposition approach can determine the high frequency Contourlet sub-bands, which contain details such as texture and edges; and the low frequency Contourlet sub-bands that contain the fuzzy, spread-out information such as background information of an image. This property of WBCT allows determining a fusing criterion where the best areas of a set of images are selected, according to the level of energy of the sub-bands. However, the investigations revealed that when

adopting above WBCT approach, the output images added undesirable effects such as loss of resolution, ghosting of moving objects, blockiness, chromatic aberrations (de-colouring around the edges), and a hazy appearance of the image due the LP decomposition/reconstruction, average blending of wavelet transform band and a non-consistent selection of areas from the low and high sub-bands (see section 5.4.2.1, above). Figure 5.6(a) shows a cropped section of a fused image obtained using the approach presented in [31]. Near the edges and around the plants in the middle of the image, blockiness, chromatic aberrations and the loss of resolution can be observed.

(a) (b)

Figure 5.6: (a) Fusion artefacts that appear in a fused image (150 % zoom) when using a WCBT fusion as in section 5.4.2.1. Visible artefacts: blockiness, loss of resolution, chromatic aberrations effect, and hazy appearance. b) Image with a ghosting effect (indicated by the arrows).

Due to the multiple images that are needed to create HDR or multi-focus images, if objects move while capturing the images, a ghosting or duplicate effect will appear in the image. Figure 5.6(b) shows a fused image with ghosting effects mentioned. The arrows in Figure 5.6(b) point to the (people walking, cars, and the lorry) ghosting effects.

### 5.4.3.1 Approach - 2

In order to solve the appearances of artefacts, haziness and to reduce the ghosting effects, in the proposed algorithm, the image only decomposed into one level of the wavelet transform. The wavelet transform produces four sub-bands, and on each high pass wavelet sub-band a directional Contourlet transform is applied. By analysing the bands, a new fusion rule is proposed for the fusion stage of the proposed WBCT approach-2. Figure 5.7 illustrates the frequency partitioning of a given image when WBCT is used. In Approach – 1 a three-level Wavelet transform and the eight-direction Contourlet transform is applied. In approach -2 first, a one-level decomposition of wavelet transform is applied on an image giving nine high pass and one low pass sub-bands. Then a three-direction Contourlet transform is applied on each high pass sub-band of the wavelet decomposition.

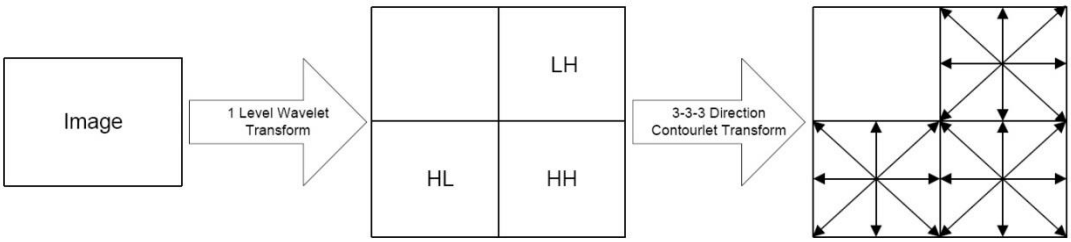



Figure 5.7: Frequency partitions obtained with WBCT

Figure 5.8 illustrates the block diagram of the fusion module. Note that for clarity of illustration Figure 5.8 does not illustrate the low-frequency Contourlet transform.

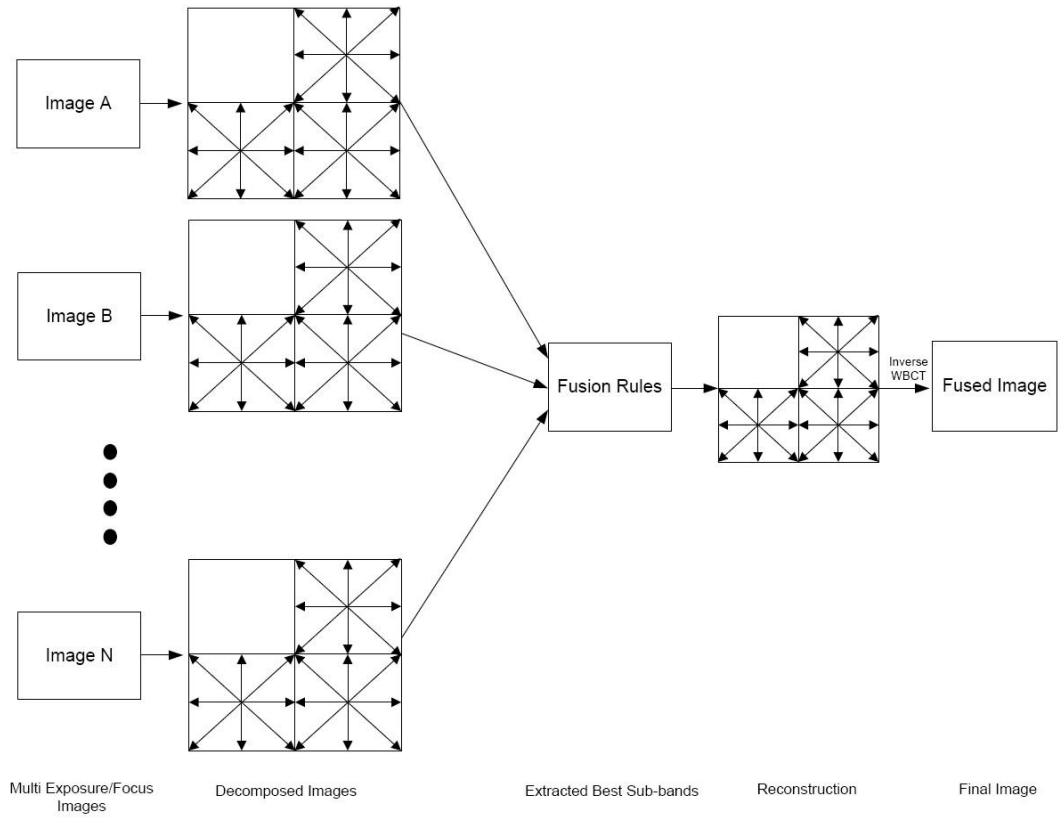



Figure 5.8: Fusion module

This fusion process is explained in the following section.

***I - Fusion of All High and Low Frequency Contourlet Sub-bands:***

The basic idea of the fusion algorithm is to generate a generic fusion mask that can be used to fuse all high, and low frequency Contourlet sub-bands, and the low pass wavelet sub-band.

In order to generate the fusion mask, the absolute energy of high-frequency coefficients is calculated in a block-based manner, comparing and selecting the

block with the higher energy (by comparing the two corresponding blocks in the under and over exposed images). The fusion mask records in a binary fashion the selected block position in the image according to the correspondence of the block selected. The energy of a region can be calculated as follow:

Region energy $E$ of a high frequency sub-band $E_H$ (where $H = (l, m, n)$, l-level of wavelet decomposition, m-LH, HL and HH bands of wavelet decomposition, n-directional Contourlet sub-band) of an image $I$ can be calculated using equation 5.6.

$$E_H^{(X)} = \sum_{(x,y)\in H} f_H^{(X)}\ (x, y)^2 \qquad 5.6$$

Where $f_H^{(X)}(x, y)$ is the coefficient at location $(x, y)$ of the high frequency sub-band $H = (l, m, n)$.

Figure 5.9 shows an example of a fusion mask obtained from the calculation of the energy of the high frequency Contourlet sub-bands. Note that the fusion mask is a binary image in which a value 1 represent the fact that the higher energy block comes from the under-exposed image and a value 0 represent the fact that the higher energy block comes from the over-exposed image. Further for illustration purposes only the fusion mask is drawn to be of similar in dimensions to the original images in Figure 5.9. It is noted that a single element of the fusion mask represents an entire block area of the original images.

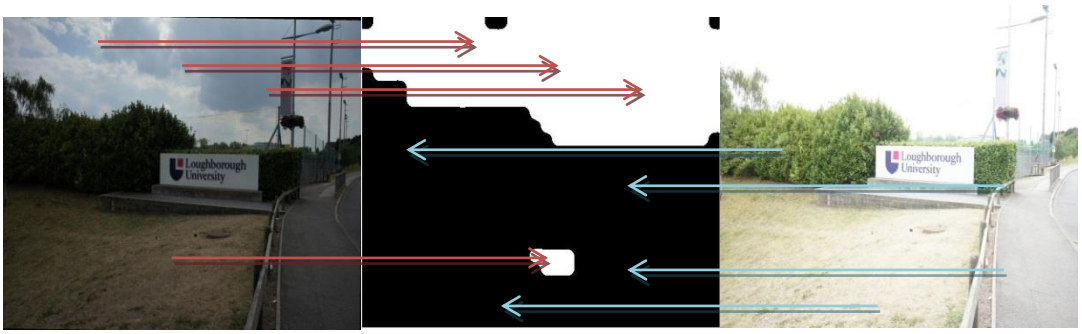


Figure 5.9: Fusion mask (middle image) obtained from the high frequency Contourlet sub-bands absolute energy of high-frequency coefficients of all three wavelet decomposition bands (LH, HL and HH)

After the fusion mask has been created, all the high (i.e. LH, HL and HH) and low frequency Contourlet sub-bands can be combined following a block by block approach, using an alpha blending approach, as follows:

$$SBout = SBu * \alpha + SBo * (1 - \alpha) \qquad 5.7$$

Where $\alpha$ is the value of fusion mask for the block, $SBu$ is the sub-band of the under exposed image, and $SBo$ correspond to the sub-band of the over exposed image. $SBout$ is the sub-band of the fused block.

Equation 5.7 works the same way when fusing multi-focus images, with the multi-exposure images replaced by the multi-focus images.

Once all Contourlet sub-bands are alpha blended, the low pass wavelet sub-band is fused as described in the next section.

***II - Fusion of Low Pass Wavelet Sub-band***

The low-pass wavelet sub-band holds the luminance information of an image. Thus, in order to generate a HDR image it is important to keep the best luminance areas of all images. i.e. clipped dark or clipped bright areas should not be considered in the fusion of the low-pass wavelet sub-band. WBCT previous fusion attempts fused the low pass wavelet sub-band by calculating the average of the low pass wavelet sub-bands of the multiple exposure images. However the investigations revealed that this approach produced a hazy fused image. In order to solve this shortcoming, the fusion mask generated by the calculation of the energy of the high frequency Contourlet sub-bands used again to fuse the low-pass wavelet sub-band. However in this case, the fusion mask is blurred with a Gaussian kernel with standard deviation of 1.3 prior to its use in fusion. After blurring the fusion mask, the low pass wavelet sub-bands are obtained using equation 5.7.

Figure 5.10(b) shows the fusion mask obtained from proposed fusion approach after being blurred with the Gaussian kernel.

(a) (b)

Figure 5.10: (a) Fusion mask obtained proposed fusion approach. (b) Fusion mask with blurred with a Gaussian kernel

### 5.4.3.2 Reconstructing the Fused Image

After obtaining the best low and high frequency sub-bands of Contourlet transform using the approach presented in section 5.4.3.1 and best low frequency sub-band of wavelet transform as described in same section, the complete fused image is reconstructed using inverse WBCT approach.

### 5.4.3.3 Experimental Results – Approach-2

All the results presented in this section were evaluated by professional image quality assessors from industry. The afore mentioned experts all concluded that the images produced by the proposed approach were superior to those which they had come across during the course of their work. Figure 5.11 represents experimental results on a set of five multi-exposure image pairs. For each set in a top-to-bottom and left-to-right order the five images illustrated represents, the under exposed image (a), the over exposed image (b), the output obtained with the WBCT approach of [31] with no compensation for camera-shake (c), the same with compensation for camera shake using the approach proposed in sections 5.3 and 5.4.2.1 (d) finally the fused image obtained when using the camera shake compensation approach and the novel WBCT approach presented in section 5.4.3.1 (e). The results clearly demonstrate the fact that the removal of camera shake is vital for a good quality fused image and can be successfully addressed by the novel approach present in section 5.4.3.1. The results also demonstrate that the use of the

proposed novel WBCT approach to fusion results in images which are sharper and more visually pleasing and appears natural.

(a) (b)

(c) (d)

(e)


Loughborough University
Loughborough University

(a) (b)


Loughborough University
Loughborough University

(c) (d)


Loughborough University

(e)

(a)

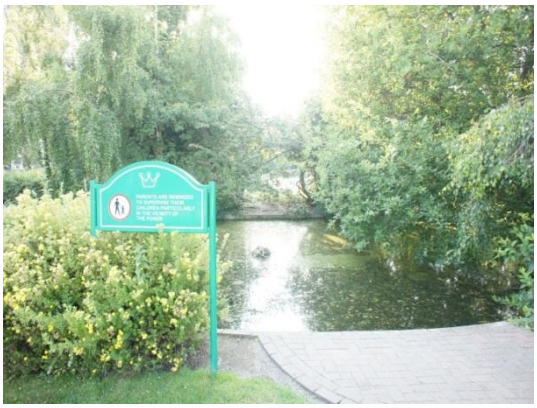
(b)

(c)

(d)

(e)

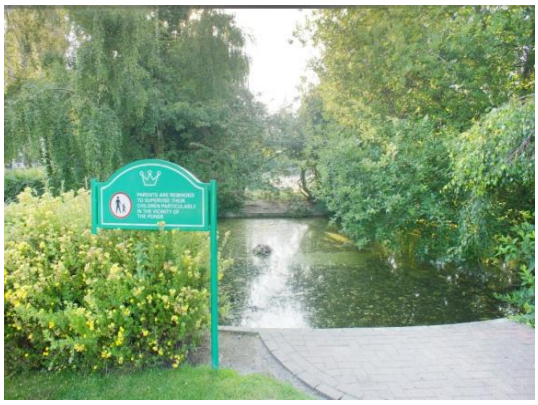
(a)



(b)

(c)

(d)



(e)

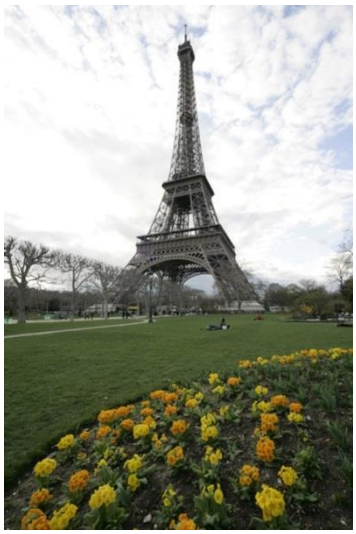
(a) (b)

(c) (d)

(e)

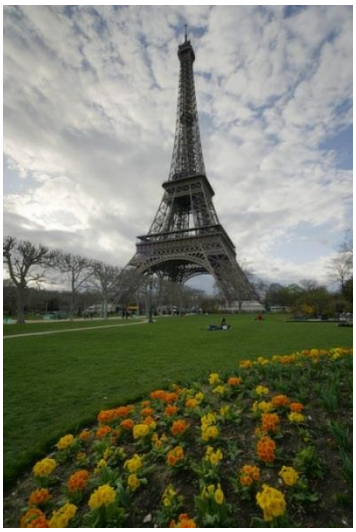
Figure 5.11: Comparing the performance of the proposed approach to two existing algorithms

Figure 5.12 compares the performance of the proposed algorithm with three state of the art algorithms namely, previous approach proposed in section 5.4.2.1 to multi focus and multi-exposure image fusion presented in section 5.4.3.1 and the outputs generated by two commercial products namely Photomatix and Photoshop. A visual comparison of results presented in Figure 5.12 demonstrate the proposed algorithms ability to produce HDR images that are visually pleasing, free of artefacts and appears to have a wide dynamic range and contrast.

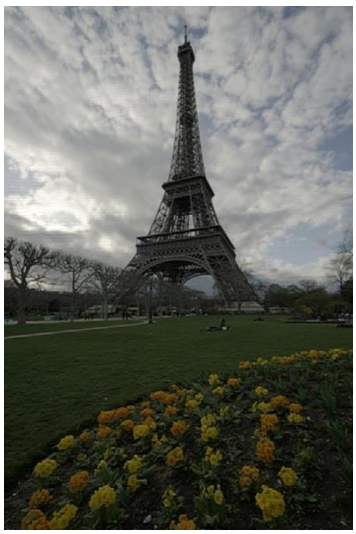


(a)

(b)

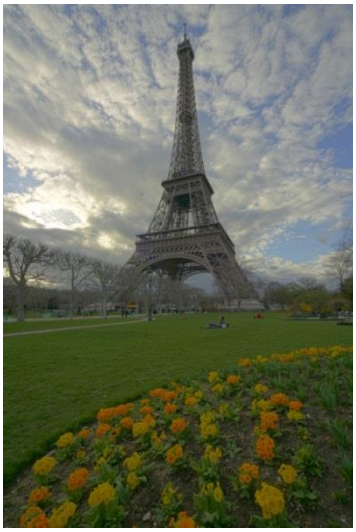


(c)

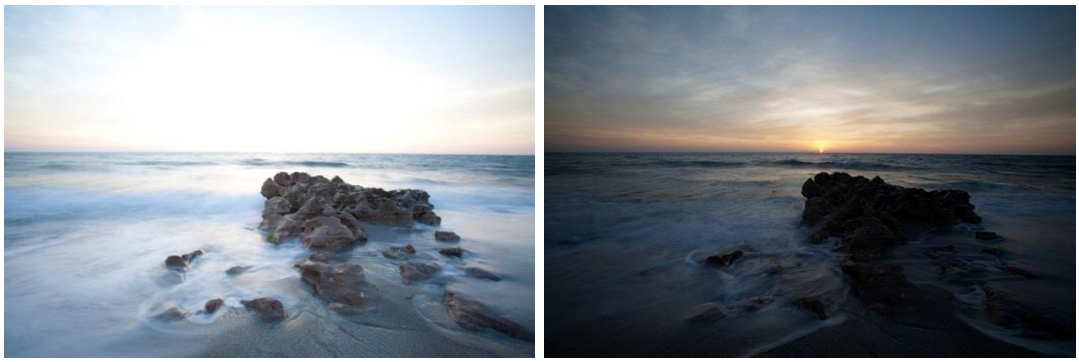


(d)

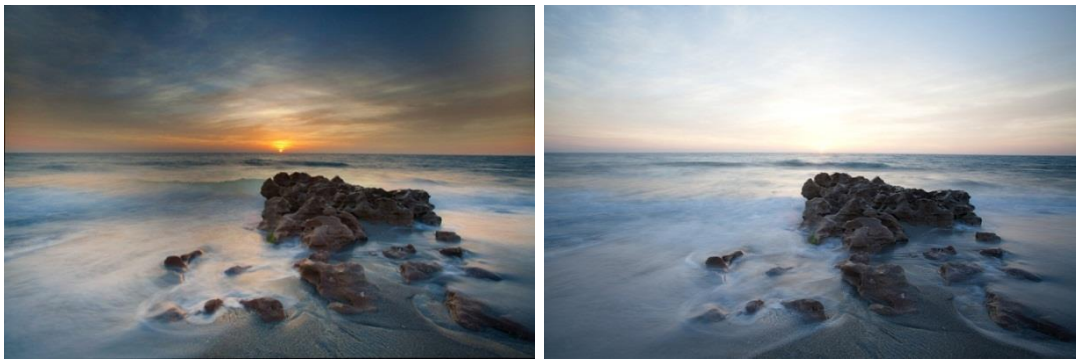
(e)



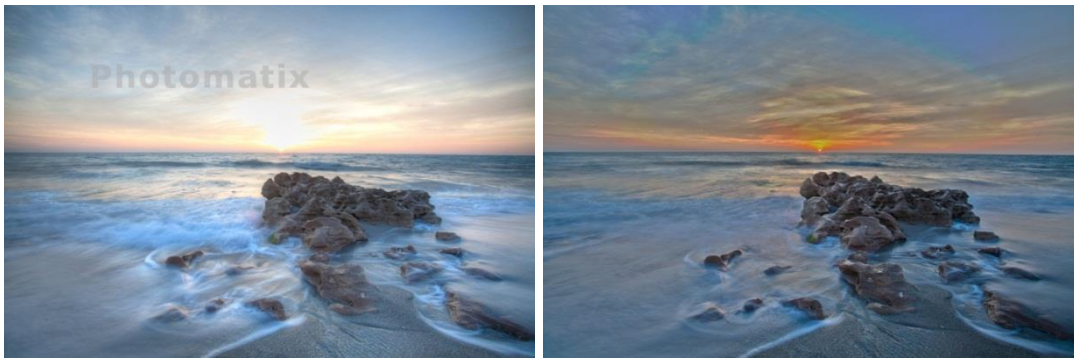
(f)



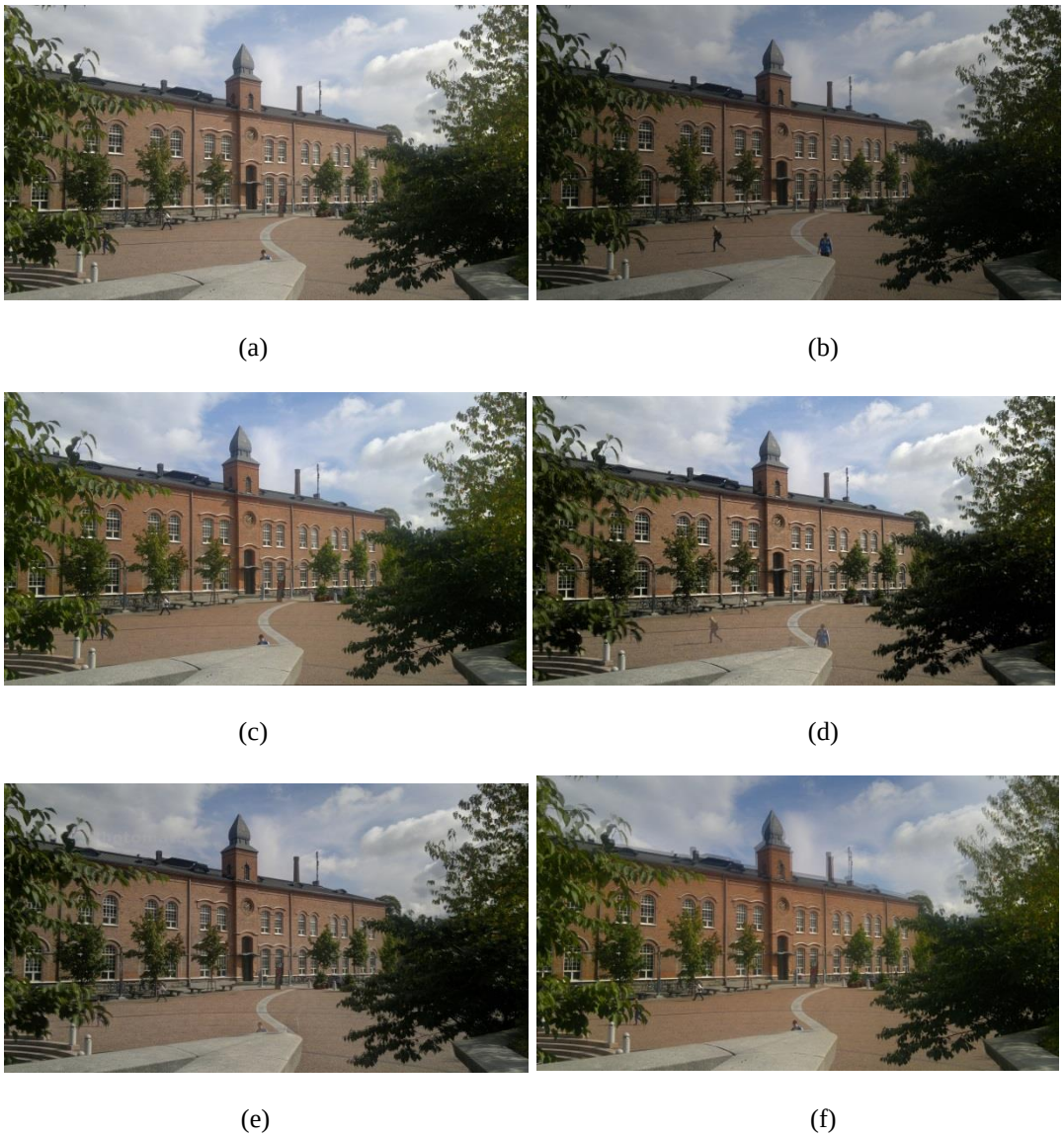
(a) (b)



(c) (d)




(e) (f)



(a) (b)

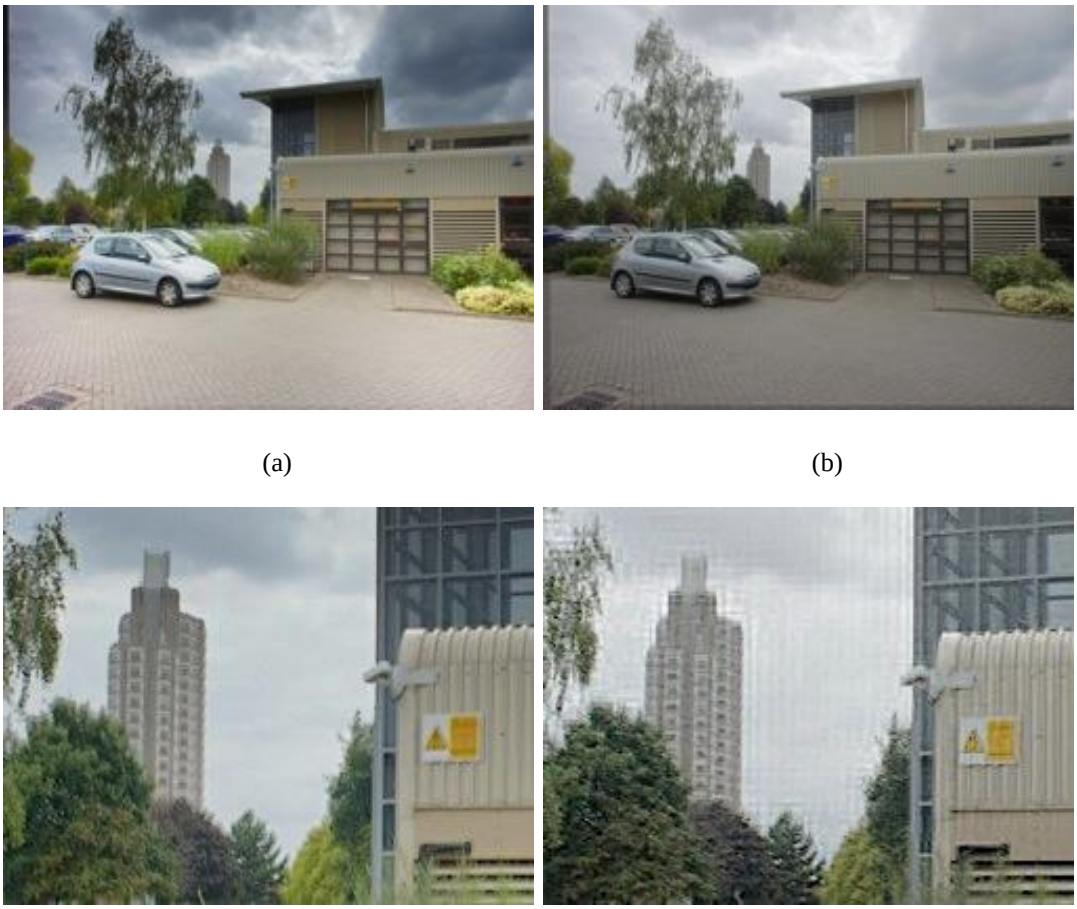
(c) (d)

(e) (f)

Figure 5.12: Comparison of proposed approach with state of the art approaches. (a) and (b) are Under exposed and Over exposed images respectively. The outputs produced by (c) the proposed approach in section 5.4.3.1, (d) approach proposed in section 5.4.2.1, (e) that of Photomatix and (f) Photoshop.

Figure 5.13 presents results of a more detailed investigation into the impact of the proposed fusion algorithm-2. The WBCT based fusion approach of [31] is used in order to compare the performance of the proposed approach. The comparison of

Figure 5.13(a) with Figure 5.13 (b) shows that the proposed approach is capable of enhancing the visible dynamic range of the image, making it look more visually pleasing. In Figure 5.13 (a), the image haziness and colours are better preserved than in Figure 5.13 (b). In order to illustrate the proposed approaches ability to enhance the sharpness of edges, colour aberrations, smudginess, and the reduction of blocking artefacts, Figure 5.13 (c), and 5.13 (d) illustrate a cropped and significantly zoomed area of the images that were illustrated in Figures 5.13 (a), and 5.13 (b). The proposed approach does not add unwanted artefacts (see Figure 5.14 (c)). It is noted that in the results being presented in Figure 5.13 all experiments were conducted on a pair of images that underwent prior registration to compensate for camera shake, using the novel approach that was presented in section 5.4.2.1. , i.e. any possibility of camera shake has been removed and the comparisons of quality is being done after this initial stage of the proposed system, the basic idea being to prove the added value of the novel approach to WBCT based fusion algorithm presented in section 5.4.3.1.



(a) (b)

(c) (d)

Figure 5.13: (a) fused image with the proposed methods. (b) Output of the previous WBCT approach. (c) Cropped and zoomed area of (a) showing improvements in colours aberrations, blockiness artefacts, and smudgy and blurry edges. (d) Cropped and zoomed area of the previous approach of the WBCT output.

When objects in the image moves at the time of capture, ghosting effects may appear in the fused image. Figure 5.14(a), shows how proposed approach is less sensible to ghosting artefacts in comparison to the previous approach (Figure 5.14 (b)) based on WBCT presented in [31].

(a)

(b)

(c)

(d)

Figure 5.14: (a) Proposed approach result, (b) Fusion artefacts that appear in a fused image (100 % zoom) when using a WCBT fusion as in [31]. Visible artefacts: blockiness, loss of resolution, chromatic aberrations effect, and hazy appearance. (c) Proposed approach result, (d) Image with a ghosting effect (indicated by the arrows).

Due to the nature of the WBCT approach adopted, the proposed algorithm could also be used to fuse a set of multi-focus images. In the experiments carried out it could be verified that the quality of the multi-focus images fused were preserved, and in most cases enhanced in comparison to other fusion algorithms such as: fusion using Laplacian Pyramid (LP) decomposition/reconstruction and spatial frequency fusion. Figure 14 shows the multi-focus images fused and comparison of the outputs of the algorithms mention above. Figure 14 shows two sets of images each marked from (a-d) that were used to evaluate the performance of the proposed approach. The results clearly show that the proposed approach performs better than the spatial frequency algorithm based and the LP fusion based approaches. It is noted that although the LP fusion approach did not add artefacts the resulting fused image is perceptually blurred (Figure 5.15(e)). The spatial frequency algorithm added stitching artefacts such as over sharpening. Further luminance wise flat backgrounds made the algorithm fail to output a visually pleasing image (see Figure 5.15(d)).

(a)

(b)

(c)

(d)

(e)


(a) (b)

(c) (d)

(e)

Figure 5.15: Multi-focus image fusion comparison: (a) and (b) show the original set of multi-focus images. (c) is the result of the image fused with the algorithm proposed. (d) is the resulting fused images using a LP fusion approach; and in (e) a spatial frequency fusion approach is used.

On the experiments performed with multi-focus images, the depths of fields had to be restricted to a certain focus range. This is because objects can be bigger in appearance from one image to another. If the range would not be restricted, occlusion problems will arise and it will make the fusion algorithms tested to fail (add occlusion artefacts) when fusing the images.

The results of the images fused with the proposed algorithm clearly illustrated how and why the proposed approach can compensate for camera shake and also showed the improvement of subjective quality performance when fusing multi-exposure and multi-focus fusion in comparison to other fusion algorithms.

## 5.5 Conclusion

In this research work a new algorithm - capable of compensating for camera shake of a number of degrees of freedom and capable of producing HDR images and effectively fused multi-focus and multi focus images - was proposed. This algorithm enables users to create HDR and multi-focused images with a SDR camera without the aid of stabilising devices such as tripods, due to the compensation in place for the removal of camera shake. The proposed approach registers a set of multi-exposure or multi-focus images with an algorithm that is based on SIFT feature point selection, followed by the use of RANSAC algorithm for removing outliers in matching, and finally a fast CPD algorithm which uses a fast Gauss transform and low-rank matrix estimation to register displaced images. After the set of multi-exposure images are registered, two novel wavelet based

Contourlet transform approaches for image fusion is used. The experiments conducted with the proposed approaches enabled us to separately demonstrate the positive impact of both the camera shake compensation algorithm and the image fusion algorithm. In particular their performance was compared with a number of state-of-the art approaches including those of the most popular commercial products. Experiments revealed the proposed approaches ability to significantly improve the visual quality of fused images, minimising or completely removing unwanted artefacts often created by other state-of-the-art approaches. Results were demonstrated using a large set of test images captured specifically for the experiments conducted.

# CHAPTER 6

# Vehicle Make and Model Recognition in CCTV Footage

## 6.1 Introduction

Several vehicle monitoring and security systems are based on the automated number plate recognition (ANPR). For example, an ANPR system could be used to prevent illegal entry into a particular location; or to enforce traffic laws or taxes in a city; and could also be used to track vehicles in the event of any crime. One of the available ways of circumventing monitoring and security systems that are based on ANPR is number plate forgery; when a person clones an already registered number plate so as to either gain entrance into a facility as the original owner of the number plate, or to evade tracking after committing a criminal offence. Due to the abovementioned number plate forgery, ANPR systems are not sufficient by themselves to ensure proper security. One way to solve this problem would be to augment existing security systems that use ANPR with a vehicle make and model recognition (VMMR) system.

A VMMR system utilizes computer vision techniques to determine the make and model of a vehicle e.g. BMW X3, FORD Focus etc. With such a system, it will be

possible to match the vehicle's plate numbers with the pre-registered make, model, and even colour of the vehicle. This will make number plate forgery a more difficult task, although not completely impossible. Other possible applications of VMMR are tracking and marketing research. The VMMR system could also provide improved surveillance and tracking in the event of crime. Usually when a crime is committed and the criminal escapes using a vehicle, police officers collect a description of the criminal and the escape vehicle. These officers now have to review large amounts of surveillance footage to find the criminal using the given vehicle description. With a VMMR system, the amount of footage to be watched could be reduced to the few places where a particular make and model has been discovered thereby speeding up the capture of the criminal.

Another use of VMMR could be marketing research; a company might decide to do a survey about the vehicles they manufacture or sell. They may want to find out how many of said vehicles are being used in a region and target their marketing and advertising based on the results of the survey. A VMMR system can provide an automated way of performing such a survey and returning very accurate results.

In much of the research conducted in VMMR, high quality static images or videos are used and the vehicles to be recognized are manually cropped out, tested, and results from this cropped datasets are provided. The goal of this project is to develop vehicle make and model system that works with low quality CCTV footage. The system should accept video data, extract the vehicles from the video, and return the make and model of the extracted vehicles. The performance of this complete system will be evaluated.

## 6.2 Proposed System

The proposed VMMR system is designed to be used with video and not still images as in much of the previous research; hence, it is designed to take advantage of the temporal resolution of videos. This system will take the feature-based approach to solving the VMMR problem. From the literature, feature-based approaches have given a minimum classification accuracy of 59%, therefore a 59% classification rate will be considered as current system's success threshold [48]; any accuracy rate above this minimum would imply proposed system can match some proposed systems. The proposed VMMR system can be divided into 5 stages. The stages are illustrated in Figure 6.1.

- Automatic License Number Plate detection
- De-skewing
- Object Tracker
- Region of Interest selection, correction, and pre-processing
- Feature Extraction
- Classification

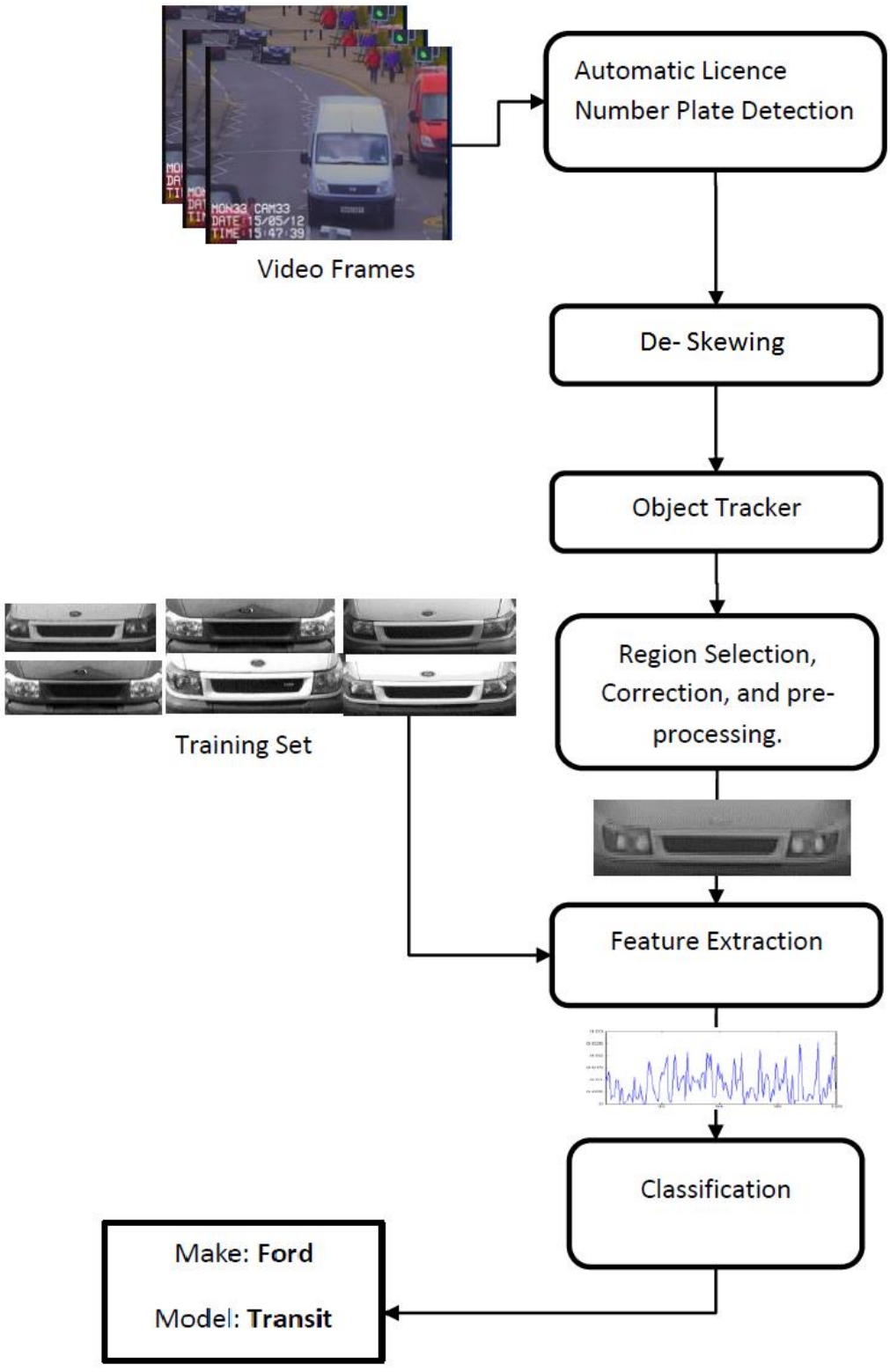



Figure 6.1: The Structure of the proposed VMMR System.

### 6.2.1 Automatic Licence Number Plate (ALNP) Detection

This stage of the VMMR system is essential because it is used for the selection of the region of interest. The segmentation of an appropriate region of interest for VMMR is a non-trivial issue due to the high variability in the design of vehicles. A solution to this segmentation problem is to select a region around a structure of reference that is common for all vehicles. The number plate is typically used as a structure of reference; It is an excellent choice as every road worthy vehicle should have a number plate, and they are mostly rectangular in shape and located at the same position on vehicles (at the centre, and near bottom) with a few exceptions.

In the Automatic Licence Number Plate (ALNP) detection stage, the image of the vehicle is processed, the frame is scanned for the shapes and margins of connected components which are then processed based on template matching (in this case it will be a rectangle) to eliminate the false ROI. Selected regions are chosen on the basis of their size and aspect ratio. If the size of region is smaller or larger than a specific threshold, it will be classed as false region and is discarded from processing. Also if the aspect ratio of any region is less than or greater than a certain threshold, then that region is also discarded for further processing. The remaining candidate regions are then checked based on their texture similarity to license plate-like areas. Then locations and the measurements of the four corners of the detected number plates are returned [77]. Each detected number plate is considered as the location of a vehicle which must be recognized. The disadvantage

of using the number plate as a reference structure is that when the number plate is occluded the vehicle will not be detected.

In the above number plate detection stage, once the algorithm locates a rectangular region, if it appears either skewed or rotated on the 2D image plane the rectangle obtained would be the one that contains the four edges of the rectangle. The lines connecting the two upper or two bottom corners of the number plate may not be horizontal, and the lines joining the two left-side and two right-side points may not be vertical (see figure 6.2(b)). Under this condition if the remaining stages of the vehicle segmentation are followed the segmented area will not symmetrically include the vehicle. This may lead to significant errors in VMMR rates. Therefore a de-skewing stage is introduced as detailed below.

### 6.2.2 De-Skewing

De-skewing is a very crucial part of the VMMR system. Due to the road gradient and position of CCTV cameras, cars coming via a curve road (bend) or lopsided road will appear at a slight angle (skewed) which will result in a wrong detection: as the front view of the car will not appear straight or symmetric. Hence, correcting the angle of a car at an early stage of the VMMR system will benefit the car detection and recognitions. The De-Skewing algorithm which is proposed to tackle this problem is explained below.

A template of a car approaching the camera at a right angle to the camera plane is initially selected. (Figure 6.2(a)). Then the width and height of the detected number plates are returned from ALNP stage.

Using the position of four corners of the number plate in video frame obtained from the ALNP detection stage (Figure 6.2 (b)), and template point positions, the two images are subsequently registered using the CPD algorithm [see Chapter 3]. CPD is based on Point Set Registration and forms links between two given sets of points to find the corresponding features and the necessary transformation of these features that will allow the points to be registered. In applying CPD to the registration of two number plates which are skewed respective to each other, the four corner points from each number plate are considered as the two point sets.

As the result of registration, all points in video frame will be rotated to resolve angular mismatch and the car which appeared to be skewed in its original stage will be de-skewed (see Figure 6.2) and be ready for processing by the subsequent stages as described in the following sections.


(a) (b)


(c)


(a) (b)


(c)

Figure 6.2: (a) Frontal view of template car, (b) Car approaching at an angle, (c) After registration, Image source [85]

### 6.2.3 Object Tracker

Object Tracking is the process of following an object through a sequence of images [67]. This stage has not been included in previous VMMR research; but it is a vital stage in the proposed VMMR system for the following reasons:

- With the object tracker, the system is able to maintain some information on each vehicle across multiple frames and not process each of the detections from every frame as a new vehicle.

- In the event that the number plate detection fails for a frame, it is possible to use previous predictions to predict the likely location of the number plate allowing continuity of detection.

- The object tracker could be used to reduce the number of false detections, if a certain number of consecutive detections are used to validate the object before it is tracked.

The inputs to the object tracker are the measurements received from the De-skewing section which are four corners of the corrected licence number plate. Using these measurements the centre location of the number plates are calculated and tracked. The object tracker is composed of two sub stages: The Data Association and Kalman Filtering [68] Stages. The data association stage enables the tracker to merge measurements from an object in a new frame at time k to the measurements from the previous frame at time k-1. Then the Linear Kalman filter is used with the assumption that the vehicles move linearly and that all measurements and noise have a Gaussian distribution. The first stage of the Kalman filter is to use previous

measurements to predict the measurement for the current time k. The next stage is to use the object's current measurement (if available) to update the predicted measurement.

### 6.2.4 Region of Interest Selection, Correction, and Pre-Processing

The Region of Interest could be any part of a vehicle that contains sufficient unique details (or features) to distinguish that vehicle's make and model from that of others. In previous work, the frontal and rear views of the vehicle have been found to have discriminating features e.g. headlamp/rear lamp, grill, and manufacturer's logo. The extraction of features from the rear view of the vehicle is not practical for a real world application because VMMR is considered as an upgrade for a security system; due to fact that the rear view of the vehicle is mostly never visible when attempting to enter a car park or other secured places, except if the vehicle reverses into the car park which is not common. The rear view could be used if VMMR is designed for surveillance but so can the frontal view. Therefore after these considerations it was decided that the frontal view would be used for analysis.

#### 6.2.4.1 Region of Interest Selection

A region of interest (ROI) is selected from the frontal view of the vehicle using the number plate as a reference point. The ROI is defined as $\boldsymbol{Width = 2.6 \times W}$, $\boldsymbol{Height = 0.32 \times Width}$, where $\boldsymbol{W}$ is the width of the detected number plate and $\boldsymbol{Width}$ and

***Height*** are the width and height of the Region of Interest respectively (Figure 6.3). The ROI is selected from the area above the number plate: nothing is selected from the area below the number plate. Figure 6.3 shows the ROI measurement. The ROI measurements are designed to encapsulate the front grill, manufacturer's logo, and headlamps. It is important to note that the assumption used for the ROI does not hold for all vehicles; if the number plate of the vehicle is not positioned at the centre of the vehicle the assumption fails - some vehicles have their number plate closer to the bottom of the vehicle. In this scenario, the ROI selection will not completely encapsulate the headlamp and manufacturer's logo. This was the case for the Peugeot 207, but the detailed grill proved sufficient for recognition. When working with vehicle categories that are generally larger (like Jeeps, minivans and trucks) or smaller than sedans, the ROI will either fail to properly select the whole vehicle or select too large a region (Figure 6.4). In this case, an attempt is made to adjust the ROI using a region correction algorithm (see section 6.2.4.2). It is also essential that the vehicle has not more than 2 degrees variation from the horizontal plane, otherwise a poor region will be selected which could cause the known vehicle to be unrecognized. Considering the fact that the de-skewing algorithm described in section 6.2.3 has been already applied this assumption can be considered to be valid in practice.


Figure 6.3: The region of interest measurements

(a) (b)

Figure 6.4: Scenarios in which the region selection might fail. (a) Ford Transit. (b) Peugeot 207.

### 6.2.4.2 Region Correction

As explained in the previous section, some of the selected regions will have to be adjusted. The method adopted to adjust the rectangle is explained in this section. In

this stage, the assumption is made that the vehicles are moving. The region correction algorithm is explained below. First step is motion segmentation, which separates moving objects from the rest of an image. For this purpose, the OpenCV implementation of a 'Motion Analysis and Object Tracking' algorithm is used [84]. In motion segmentation, using the current frame and the previous frame(s), a background representation model is initially learned. Subsequently the difference of the upcoming frames from the model is calculated. Finally noise is removed and any significant size region is marked as a moving part.

Then by using the measurements and location of the region selected in the ROI, if the segmented object obtained from first step is wider or smaller than the region selected, the difference (error) is computed. This difference is used to adjust (increase or decrease) the width of the region selected and a new height is computed as: $\boldsymbol{Height = 0.32 \times Width}$

Motion segmentation fails when the object is stationary. This will also cause the region correction algorithm to reduce the region excessively. Finally as a solution to this a threshold is defined; if the width is reduced below this threshold, it is assumed the motion segmentation has failed and the original region of interest is used.

### 6.2.4.3 Pre-Processing

After the region correction, the Region of Interest is cropped out automatically, the cropped out image is resized to $64 \times 152$ pixels (i.e. normalised with linear interpolation), and the same pre-processing is manually preformed on all images in

the training set which is used in the classification stage. This cropped image is converted to a grayscale image and is passed onto the feature extraction module.

### 6.2.5 Feature Extraction

Feature Extraction is a very important stage in any VMMR system. Features need to be distinctive enough to distinguish between a large number of vehicle makes and models; and at the same time, they also need to be invariant to changes in scale, colour, and illumination. These features should be able to model the shape of the local components (grill, headlamps) in the region of interest. The proposed system makes use of the Local Energy Shape Histogram (LESH) first introduced by [46], which performed satisfactorily when compared with other feature descriptors. The LESH feature vector is extracted from the region of interest (with 512 features), and normalized between [0, 1]. Theoretical aspect of LESH is explained in background chapter (Chapter 3).

### 6.2.6 Classification

In this stage, features extracted from the region of interest are passed to a classifier which would determine the make and model. The support vector machine is a classifier which has proven very reliable in previous VMMR research; therefore it has been chosen for this system. The proposed system uses multiple frames from the video in the classification stage. When a vehicle is detected, features are

extracted from them and classified; the resulting label is stored. This is repeated on a pre-specified number of frames for the same object; after which all the predicted labels for that object are used in a Winner-Takes-All voting and the vehicle is assigned the label with the majority of the votes. A random label is selected in the event of any ties.

#### 6.2.6.1 Training the SVM

When training a SVM, the first step is to pre-process and crop the training images in a similar manner as described in section 6.2.4. Then LESH features are extracted from every image in the dataset and scaled appropriately.

The next step is the selection of a kernel function for the SVM. A number of kernel functions exist; [78] recommend the RBF kernel as the reasonable first choice when selecting a kernel function. It is stated that the linear kernel is a special case of the RBF kernel and also the sigmoid kernel would perform like a RBF kernel with certain parameters. Since the RBF is a less complex kernel function when compared to the polynomial kernel, it was used first to train the SVM. Having performed optimally within the proposed system the RBF kernel function ( $k(x, y) = e^{-\gamma\|x-y\|^2}$ ) was selected as the kernel function in the final implementation. Next stage was parameter selection. Different kernel functions have different parameter values and it is thus important that the right parameter values are selected to obtain the best possible accuracy. By performing grid search on parameter $\gamma$ and $C$ using cross-validation, best measures were determined. The parameters $\gamma$ and $C$ for the

RBF kernel were selected using k-fold ($k$ = 5) cross validation. The SVM is trained using the selected parameters and a data structure is generated which contains the parameters, support vectors, and other important data. The One versus All SVM is used in this implementation because it provides the best performance and it also outputs the decision values which it used to assign labels as opposed to the One versus One SVM, which uses votes to assign a label. A SVM will assign an input feature to its closest matching class but in a real world VMMR system this is undesirable because an unknown vehicle could be classified as any vehicle in the dataset. As a solution to this each class from the training set is assigned a threshold; when an unknown feature is classified, and the decision values returned from the SVM are below the threshold for that class, the feature is re-classified as an unknown vehicle. The decision thresholds are selected empirically.

## 6.3 Experimental Results and Analysis

A number of experiments were performed to test and evaluate the performance of the implemented VMMR System. The results from each of these experiments are presented and discussed in this section.

### 6.3.1 Vehicle Database

The vehicle database used for this project consists of a training dataset of static images and the testing dataset of videos. These are discussed briefly below.

### 6.3.1.1 Training Dataset

The training dataset used comprised of frontal views of 22 different vehicle makes, with 7-10 images per make and model, making a total of 196 images (Figure 6.5 - summary is in Table 6.1). The database is created from images sourced from vehicle auction and sale websites, vehicle review blogs, and some images provided at the start of the project. Most of the images are taken in bright and clear conditions with about 5 degrees variation from the horizontal plane; with a few exceptions which were adjusted in Photoshop. The database contains some difficult types like the Renault Clio campus sport and Renault Megane which both have similar grill and headlights. The database also contains a few trucks and a jeep e.g. BMW X3, Ford Transit etc. The images collected have varying resolutions but all the images are cropped and normalized using the same procedure described in the region selection and pre-processing sections in section 6.2.4.

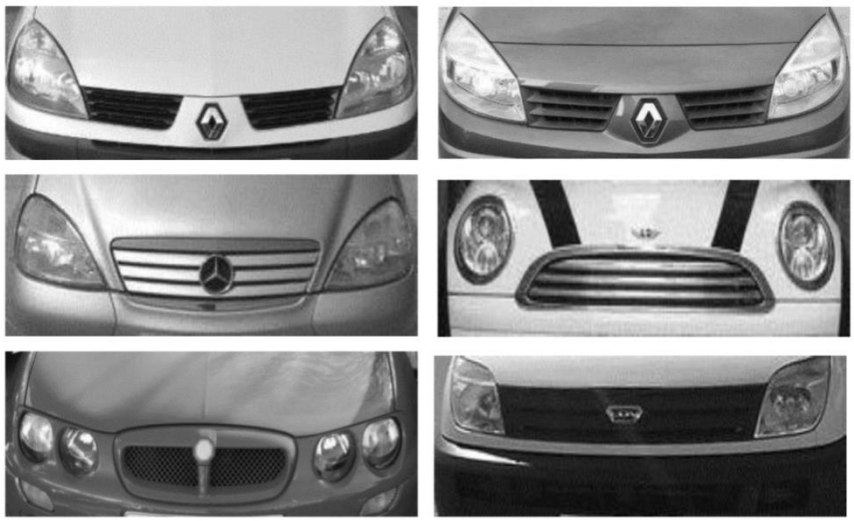


Figure 6.5: Samples of the images from the Training dataset showing 6 different vehicle make and models.

| S/n | Make | Model | No of Sample | S/n | Make | Model | No of Sample |
|---|---|---|---|---|---|---|---|
| **1** | BMW | X3(2006) | 9 | **12** | Mercedes | Sprinter | 8 |
| **2** | Citroën | Xsara(1997) | 7 | **13** | MG | ZR+ | 8 |
| **3** | Ford | Transit | 7 | **14** | Mini | Cooper | 9 |
| **4** | Ford | Fiesta(1999) | 10 | **15** | Peugeot | 207 Sport | 9 |
| **5** | Ford | Focus(2002) | 10 | **16** | Renault | Clio Sport | 10 |
| **6** | Ford | Focus (2004-2006) | 10 | **17** | Renault | Megane | 10 |
| **7** | Ford | Mondeo (2005) | 10 | **18** | Toyota | Yaris (2012) | 10 |
| **8** | Honda | Civic-Type-S | 9 | **19** | Vauxhall | Astra | 8 |
| **9** | Honda | Jazz(2004) | 8 | **20** | Vauxhall | Corsa | 10 |
| **10** | LDV | Maxus | 9 | **21** | Vauxhall | Meriva Breeze | 8 |
| **11** | Mercedes | A140 | 8 | **22** | Volkswagen | Tr-Porter | 9 |

Table 6.1: A summary of the Training dataset showing make, model, and number of training samples.

### 6.3.1.2 Testing Dataset

The Testing dataset comprises of 3 video clips designated as TestVideo1, 2 and 3. Together these videos contain 24 vehicles that have the same make and model as the ones in the training dataset and also some unknown vehicles The clips were taken with a CCTV camera, with a resolution of 720 x 576 pixels, frame rate of 25fps, duration of each video is between 1-2 minutes, and all the videos are stored in AVI file format (figure 6.6).

The testing dataset contains some of the challenges that a real world VMMR System might encounter. For example vehicles with their headlamps switched on, a vehicle which had stickers on the bonnet, and vehicle body or headlamps reflecting the sunlight.


(a) (b)


(c) (d)

Figure 6.6: Samples of the video frames from the testing dataset showing 4 different vehicle make and models. (a) A vehicle higher than a normal vehicle. (b) A normal vehicle. (c) A normal size vehicle with the stickers on the bonnet. (d) A higher than normal vehicle with stickers on the bonnet.

### 6.3.2 Experiments

A number of experiments have been carried out to evaluate the proposed system; the results of these experiments are provided in the section that follows. All experiments are conducted using the testing dataset described in section 6.3.1.2. To reduce the computational cost of running the system, a search area is defined; vehicles will be recognized only within that search area. It is important to note the following about selection of a search area:

- Results could differ depending on the location of the search area selected.
- The further the search area is from the camera, the poorer the recognition rate, this because at some point the scale of the vehicles will be too small.

With this in mind, each experiment is performed using a fixed search area. The measurement of this search area is specified as the default search area in the implemented system, this search area will be used if no search area is manually specified, and this is done so each experiment can be replicated. The location of the search area is defined using the x and y coordinates of the top left edge of the search area (311, 420) and the x and y coordinates of the bottom right edge of the search area (720, 520). The number of frames used for the classification is set at five (5) for all experiments except otherwise stated. Also the decision thresholds used for classification (Section 6.2.6), is deactivated for all experiments. This will allow simplified presentation of the result without the inclusion of the unknown vehicle class.

The accuracy of each experiment is calculated as the ratio of the number of correct recognition to the total number of vehicles in the test dataset.

$$Accuracy = \frac{Number\ of\ Correct\ Recognition}{Total\ number\ of\ vehicles}$$

#### 6.3.2.1 Experimental Results

The performance of the proposed system is presented in this section. The proposed system achieved a classification accuracy of 95.83%; and correct recognition on all classes except the Mercedes A140 (in the video the Mercedes A140 stops as soon as it enters the search area). The region correction fails since the vehicle is stationary and it is classified as a Vauxhall Astra. Although the experiments are performed on

a relative small number of cars, it is important to note that most VMMR experiments are performed using high resolution images but this experiments are performed using low resolution CCTV video and still acquired high recognition rates. The results are presented using a confusion matrix shown in Figure 6.7, where the class number corresponds with the classes in Table 6.1.

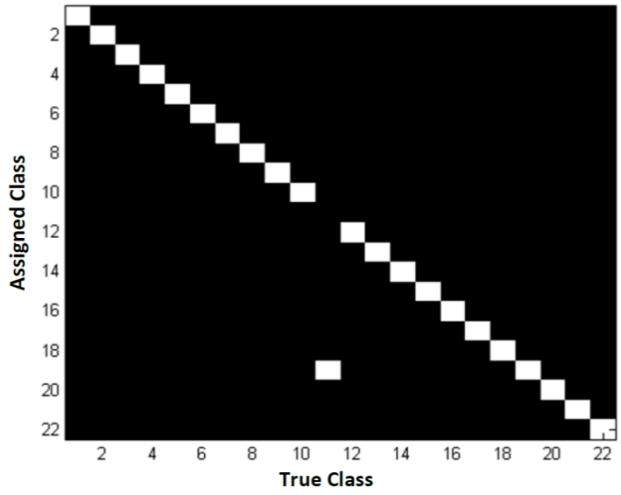



Figure 6.7: Confusion Matrix showing the classification results for each class.

### 6.3.2.2 Alternate Features and Classifiers

In these experiments, the performance of 3 features vectors is compared against the proposed feature vector, and 4 classifiers are also tested. These experiments are done to check if a better performing feature extraction method or classifier could be selected. The features compared in this experiments are: 1) Sobel Response 2) Histogram of gradients 3) LESH-128 (4x4 partitions) 4) LESH-512 (8x8 partitions). The classifiers used in this experiment are: 1) KNN (K=1) 2) KNN (K=3) 3) SVM (One versus One) 4) SVM (One versus All). Classification of new samples for the

one versus all case is done by a winner-takes-all strategy, in which the classifier with the highest output function assigns the class. For the one versus one approach, classification is done by a max-wins voting strategy, in which every classifier assigns the instance to one of the two classes, then the vote for the assigned class is increased by one vote, and finally the class with the most votes determines the instance classification [86]. The classification rates for each feature and classifier combination is presented in the Table 6.2 and Figure 6.8.

| | **LESH-128** | **LESH-512** | **HOG** | **Sobel Response** |
|---|---|---|---|---|
| **KNN (K=1)** | 70.83 | 70.83 | 62.50 | 25.00 |
| **KNN (K=3)** | 70.83 | 70.83 | 62.50 | 33.33 |
| **One Versus One SVM** | 70.83 | 91.67 | 58.33 | 54.17 |
| **One Versus All SVM** | 75.00 | 95.83 | 66.67 | 62.50 |

Table 6.2: Classification rate of different features using multiple classifiers.

The theory behind LESH has been discussed in Chapter 3. The original LESH (LESH-128) uses 4x4 partitions generating a 128-dimensional vector, and the modified LESH (LESH-512) is generated using 8x8 partitions generating a 512-dimensional vector. Histogram of gradients (HOG) [79] is a feature descriptor originally designed for pedestrian detection, but since development has been used in other object recognition tasks. It is invariant to both geometric transformation and

changes in illumination. The implementation written by [81] and provided in [80] was used. The HOG features are generated over 64 rectangular cells and using 9 bins thereby making a 576-dimensiional feature vector. The Sobel response was extracted by convolving the region of interest with both the horizontal and vertical Sobel convolution kernels, and then calculating the magnitude. The magnitude at each pixel in the ROI was used as the feature vector.

The worst performing feature vector was the Sobel response; it returned the lowest classification rate in comparison with the other features irrespective of the classifier used. This was expected as gradient based edge detectors are typically sensitive to noise [62]. The Sobel response would only be suitable in controlled conditions.

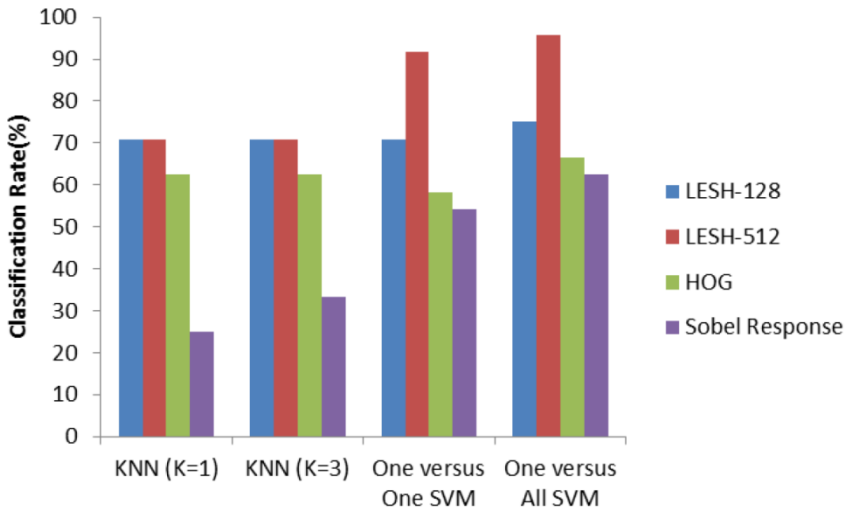



Figure 6.8: Performance comparison of different features using multiple classifiers.

The best classification accuracy is obtained by using the proposed feature vector (LESH-512) with the proposed classifier (One versus All SVM). This is to be expected because the LESH uses the local energy model which is able to extract

edges, contours, and corners at the same time thus making it a very robust feature and using 8x8 partitions makes LESH-512 more discriminating. The One versus One SVM perform closely to the One versus All SVM and could outperform it if the number of classes is increased significantly.

It is important to note that the results of these experiments could vary considerable using another dataset. The tested features like HOG have a number of parameters that could also influence the results. A few parameters were tested and the best results are provided.

#### 6.3.2.3 Number of Frames for Classification

The experiments are aimed at finding the optimal number of frames for the classification process. It should be mentioned that using smaller or larger number of frames could lead to misclassifications. It is also possible that the vehicle could leave the search area without getting classified if a very large number is specified.

For the testing dataset used, 1, 5, or 7 Frames gave the best accuracy. At 13 Frames, some vehicles began to exit the search area unclassified causing the rate of recognition to decline. For a real world system, the use of a single frame for classification might be less accurate than using 5 or 7 frames but using the available dataset it works well (Table 6.3, Figure 6.9).

| Number of Frames | Classification Rate (%) |
|---|---|
| 1 | 95.83 |
| 3 | 91.67 |
| 5 | 95.83 |
| 7 | 95.83 |
| 13 | 91.67 |

Table 6.3: Classification results from the proposed system using varied number of frame(s) for classification.

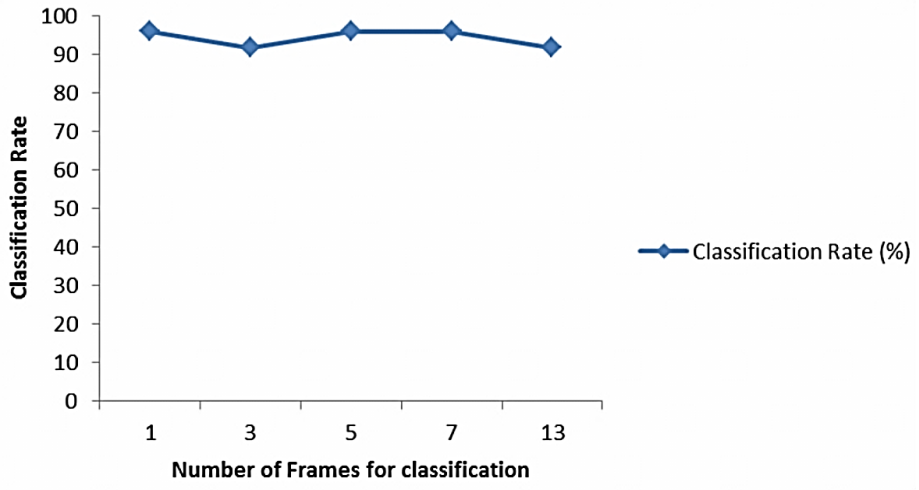



Figure 6.9: Classification results from the proposed system using varied number of frame(s) for classification.

It should be noted that the number of frames that is used to make the best judgment in VMMR will depend on many practical factors affecting the video capture setup, quality of footage, the path taken by the vehicle, size of vehicle, location of analysis window etc. However the above experiments reveal that VMMR in video, attempted for the first time, within the research context of this thesis, provides a number of additional advantages as compared to processing single frames to make judgement.

#### 6.3.2.4 Effect of Region Correction

The region correction algorithm implemented used approximate median background subtraction, which is a complex algorithm in terms of the memory expended. In these experiments, the region correction module is deactivated and the system is tested. The significance of these experiments is to show experimentally the importance of the region correction stage.

The region correction was deactivated from the settings dialog and the testing dataset was run through the system. The system achieved a 66.67% without the region correction having 8 vehicles wrongly recognized amongst these were the BMW X3, LDV Maxus, Ford Transit, Vauxhall Astra, Renault Megane, Ford Focus (2002), Mercedes Sprinter, and Mercedes A140. It is expected that all trucks and jeeps in the testing dataset are misclassified because a smaller region of interest was selected but the Volkswagen Tr-porter is still classified correctly. The ALNP detection implementation sometimes fails to return the correct width of the number

plate which is essential for the region selection; this leads to a badly selected region and eventually leads to the smaller vehicles being misclassified (Figure 6.10).

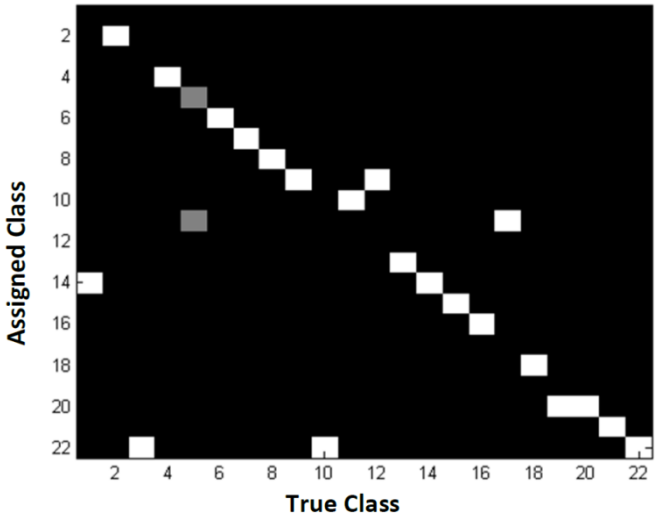



Figure 6.10: Confusion Matrix of the classification results for each class, when region correction is turned off.

These experiments proves that the region correction is significant not just for larger vehicles but for some of the smaller vehicles also.

### 6.3.2.5 Real Time Performance

In these experiments, performance of the proposed system in real time is tested. This is essential for a real world system. Although, the system is implemented in MATLAB which is typically slow, it should still use a relatively small amount of time per frame. This would prove that when the system is re-written in a language that can run quicker, it could achieve real-time speed.

The difference between the start and end of processing gives the time it took to process that frame.

The proposed system used an average of 1.011 seconds per frame when a vehicle was in the search area. This is far from the desired 25 frames per second required for real-time performance. It is important to note that the average time gained from these experiments would vary depending on the CPU speed and memory of the computer used to run the experiments. It is further worth mentioning that most CCTV cameras capture at 5 fps and processing a frame within 0.2 S is thus the ultimate goal towards achieving real time performance.

## 6.4 Conclusion

Vehicle Make and Model Recognition is a developing area of research with multiple applications. It is mainly investigated as a potential upgrade to automatic number plate recognition systems used in vehicle monitoring and security systems, due to number plate cloning becoming more common. It could also be applied in vehicle marketing research and in more robust vehicle tracking.

In this chapter, a complete VMMR System has been presented. The proposed VMMR system accepts a video and returns the makes and model of the vehicles detected in that video. The VMMR system began by automatically detecting a number plate from a video frame, the location of the number plate was used subsequently to select a region of interest. It was important to treat the same vehicle across multiple frames as a single vehicle; hence an object tracker was used to track

the detected number plate across multiple frames. A static region of interest was selected above the number plate and motion segmentation was used to adjust this static region if necessary. LESH Features were extracted from the region of interest and a SVM was used to classify the feature, assigning it a label. The final label was decided using multiple frames from one vehicle object.

A number of experiments were carried out on the proposed system using a test set that comprised of low quality CCTV videos. The proposed system achieved a best classification rate of 95.83%. Although the proposed system shows promising performance much can be done to improve the system. Some proposals for future improvement of the work presented in this chapter are detailed in Chapter 7.

# CHAPTER 7

# Conclusions and Further Work

## 7.1 Introduction

This chapter summarizes the key ideas presented in chapters 4, 5 and 6, draws conclusions and emphasizes the important contributions made by the research presented in this thesis. It also gives an insight into possible future directions of research, particularly with the intention of further extending the functionality and efficiency of the proposed algorithms.

The main motivation to the research presented in thesis came from the observation that Coherent Point Drift (CPD) approach, a relatively new fundamental theory about rigid and none-rigid transformation of point sets, can be successfully applied to a number of Computer Vision application domains to resolve some of the existing open research problems. Through this thesis, solutions to open research problems/challenges in speaker identification in teleconferencing, handling global camera shake in multi-focus and multi-exposure image fusion and vehicle make and model recognition in CCTV footage, have been presented. Further in all above application domains making the best use of CPD to address image registration, contributions were made in furthering the state-of-art in the said application domain by proposing more efficient approaches.

The presentation of this thesis was organised such that each of the contributory chapters 4-6, is revisited as a separate section, with a brief overview highlighting the motivation, novelty and contribution of that work, an extensive analysis portraying the adequacies and limitations of the proposed technique and finally making specific conclusion related to the proposed approach. These chapters lead to the following conclusions.

## 7.2 Conclusion of the Thesis

In summary, the thesis presents four original contributions to the state-of-the-art in three different computer vision application areas.

The first novel contribution presented in Chapter 4, is the application of CPD with none rigid transformations, to speaker identification in multi-people, multi-venue, video conferencing. CPD, originally proposed as a theory of point set registration provides the solution to identifying continuous non-rigid shape changes that occur in human lips, when talking. In order to be able to most effectively use CPD in this area, the Chapter-4 presented improvement to the standard face detection approach used to remove false positives and to cater for real-time needs of teleconferencing by proposing a simplified approach to lip boundary detection. Using CPD with none-rigid transformations, the lip boundary points of a detected face on each frame of the video was compared to the boundary points of a template of a closed mouth and after calculation of variance of correspondence distance using CPD, a decision was made as to whether the lips are moving or not. Experimental results on a

number of teleconferencing videos were presented to evaluate the effectiveness of the proposed system in particular for successful speaker identification.

Chapter 5 presented the second novel algorithm which was developed based on CPD, i.e. fusion of multi-exposure and multi-focus images in the presence of camera shake. To identify unique feature points in the constituent images and subsequently to determine the corresponding points that are registered using CPD (with non-rigid transformations) to compensate for camera shake, SIFT feature points (see chapter 3) were used. The general motivation behind using SIFT feature descriptors in image fusion and camera shake removal are their rotation, scale, occlusion and illumination invariance properties. This process is followed by the application of RANSAC to potentially removing the unwanted feature point outliers. Subsequently the CPD algorithm is applied to register the images to each other. As for the subsequent image fusion stage, a novel approach was proposed based on Wavelet based Contourlet Transforms (WBCT). Two variants were proposed for fusion rules after WBCT decomposition stage. The first was based on region energy, maximum variance and averaging and for second approach was based on region energy, masking based on Alpha Blending followed by the use of a Gaussian Kernel for de-blurring. The experimental results indicate that the proposed algorithm performs very efficiently and successfully. The results have been compared to two previously proposed academic approaches and also to tools provided by popular commercial software, proving the improved subjective performance demonstrated by the proposed approach.

Finally chapter 6 provided a novel approach to Vehicle Make & Model Recognition (VMMR) in CCTV footage, by using CPD to register images and thereby

compensate for skew present in the images due to camera positioning. It is noted that in a typical situation where a CCTV camera not specifically set up for VMMR, picks up the image of a vehicle, it is highly likely to be positioned at an angle to the patch of the vehicle that causes skew. This can distort the image of the vehicle and hence may result in failure of recognition of the correct vehicle make and model. It was shown that CPD can be used to provide an effective and efficient solution to this problem. The CPD based de-skewing stage was followed by an object tracking stage and then another processing corrected the ROI for the subsequent recognition stage. In this stage a new ROI selection approach was proposed based on the size of license plate to correct the problem that occurred when the height of the incoming cars were higher than normal (e.g. Lorries). Subsequently the LESH feature extraction was applied and finally using SVM the classification was conducted. Experiments were conducted on real CCTV camera videos. The results concluded that the success rate for VMMR was initially 66.67% without using the proposed enhancements, but thereafter, application of the above mentioned enhancements resulted in a 95.83% rate of success.

## 7.3 Future Work

Although a number of novel contributions to the state-of-the-art in different computer vision applications have been proposed in this thesis, it is possible to extend the work to further improve success rates and functionality of the algorithms.

All implementations of the proposed algorithms have been carried out using MATLAB. One desirable task is to implement the proposed systems in languages such as C or C++ that will increase speed up the process to real-time rates. The conversion of the MATLAB implementation to C/C++ has the ability to increase the speed and to use them within systems supported by a wide range of software platforms and hardware systems.

In chapter 4, the algorithm proposed for speaker identification is a not fully real-time as it uses Canny edge detection to identify lip boundaries which is a slow approach and can fail in some cases e.g. skin color. It is suggested that further research be carried out in identifying an edge detector that is robust to illumination change making the proposed approach real-time. Further the real time operation can be ensured by looking at ways to make use of multi-core technology and implementation of the algorithms using threads.

In chapter 5, although the use of SIFT descriptors addresses the rotation and scale invariance problems, it is not fully invariant to illumination. In order to improve illumination invariance, a number of local feature descriptors such as edge, corner, shape and texture descriptors can be combined to the SIFT descriptor used in this work. This could also be improved by use of illumination/highlight removal algorithms as a preprocessing stage. It is noted that the use of CPD with non-rigid transformations allows for compensation of camera shake both within the camera image plane and tangential to it. The experiments we have conducted do not test the algorithm's robustness to out of plane camera rotation. These experiments can be conducted and used to further justify the added advantages provided by the use of CPD.

In chapter 6 the training image database used consists of only the frontal view images and is comparatively small in size. Having multiple views allow the design and development of VMMR techniques which are based on multiple angle views of a vehicle. Having a larger database with more samples per make-model will improve the accuracy of recognition considerably due to the possibility of improving the training process. It is further noted that this algorithm was developed specifically to deal with performing VMMR in CCTV footage which is often of medium to poor image quality. Our experiments were conducted on such videos. However if the footage can be captured by video cameras recording at higher quality it would have been possible to obtain a better accuracy rates. It is noted that all current work presented in relevant literature have only being tested in purpose captured video footage.

# References

[1] M. Tistarelli, E. Grosso. Active vision-based face authentication. *IEEE International Conference on Multimedia and Expo*, vol. 4(18), pp. 299-314 ICME, 2001.

[2] M. J. Jones, P. Viola. Face Recognition Using Boosted Local Features. *IEEE Proceedings of International Conference on Computer Vision (ICCV)*, 2003.

[3] P. Yang, S. Shan, W. Gao, Li, S.Z., D. Zhang. Face Recognition using Ada-Boosted Gabor Features. *Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 356- 361, 2004.

[4] Y. Freund, R. Schapire. A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*, vol. 14(5), pp. 771–780, Sep. 1999.

[5] Y. Freund, R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, vol. 55(1), pp. 119-139, 1997.

[6] M. Bicego, A. Lagorio, E. Grosso, M. Tistarelli. Use of SIFT Features for Face Authentication. *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, pp. 35-41, 2006.

[7] B. Heisele, P. Ho, T. Poggio. Face Recognition with Support Vector Machines: Global versus Component-based Approach. *In proceedings Eighth IEEE International Conference on Computer Vision*, vol. 2, pp. 688-694, 2001.

[8] C. Cortes, V. Vapnik, Support-Vector Networks. *Machine Learning*, vol. 20(3), pp. 273-297, 1995.

[9] J. Luettin, N. Thacker, S. Beet. Statistical lip modelling for visual speech recognition. *Proceedings of the 8th European Signal Processing Conference (EUSIPCO'96)*, vol. I, pp. 147-140, 1996.

[10] P. Dalka, A. Czyzewski. Lip movement and gesture recognition for a multimodal human-computer interface. *International Multi-conference on Computer Science and Information Technology*, pp. 451-455, 2009.

[11] F. Haider, S. Al-Moubayed. Towards speaker detection using lips movements for human-machine multiparty dialogue. *In proceedings of Fonetik*, 2012.

[12] M. Bendris, D. Charlet, G. Chollet. Lip activity detection for talking faces classification in TV-content. *International Conference on Machine Vision*, 2010.

[13] S. Siatras, N. Nikolaidis, M. Krinidis, I. Pitas. Visual Lip Activity Detection and Speaker Detection Using Mouth Region Intensities. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19(1), pp.133-137, 2009.

[14] T. J. Hazen, E. Weinstein, R. Kabir, A. Park, B. Heisele. Multi-modal face and speaker identification on a handheld device. *In proceedings of the Workshop on Multimodal User Authentication*, pp. 120-132, 2003.

[15] D. J. Shiell, L. H. Terry, P. S. Aleksic, A. K. Katsaggelos. An Automated System for Visual Biometrics. *Forty-Fifth Annual Allerton Conference on Communication, Control, and Computing*, 2007.

[16] H. B. Kekre, V. Kulkarni. Speaker Identification by using Vector Quantization. *International Journal of Engineering Science and Technology*, vol. 2(5), pp. 1325-1331, 2010.

[17] D. Pullella, M. Kuhne, R. Togneri. Robust speaker identification using combined feature selection and missing data recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing,* pp. 4833-4836, 2008.

[18] J. Ching Wang, Ch. Hsien Yang, J. Fa Wang, Hsiao-Ping Lee. Robust Speaker Identification and Verification. *IEEE Computational Intelligence Magazine,* vol. 2(2), pp. 52-59, 2007.

[19] S. Kwoon, S.Narayanan, Robust speaker identification based on selective use of feature vectors. *Pattern Recognition Letters*, vol. 28(1), pp. 85-89, 2007.

[20] J. An, S. H. Lee, J. G. Kuk, N. I. Cho. A multi-exposure image fusion algorithm without ghost effect. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp. 1565-1568, 2011.

[21] I. Zafar, E. A. Edirisinghe, H. E. Bez, Multi-exposure & multi-focus image fusion in transform domain. *IET International Conference on Visual Information Engineering,* pp. 606-611, 2006.

[22] K. Kotwal, S. Chaudhuri. An optimization-based approach to fusion of multi-exposure, low dynamic range images. *Proceedings of the 14th International Conference on Information Fusion,* pp. 1-7, 2011.

[23] S. Lee, H. Wey, S. Lee. Image registration for multi-exposed HDRI and motion Deblurring. *Proceedings of SPIE Computational Imaging VII,* vol. 7246, pp. 72460-72460, 2009.

[24] M. Qiguang, W. Baoshu. The Contourlet Transform for Image Fusion. *Proceedings of SPIE Multi-sensor, Multisource Information Fusion, Architectures, Algorithms, and Applications,* vol. 6242, pp. 1-8, 2006.

[25] A. Goshtasby. Fusion of Multi-Exposure Images. *Image and Vision Computing,* vol. 23, pp. 611-618, 2005.

[26] M. Block, M. Schaubert, F. Wiesel, R. Rojas. Multi-Exposure Document Fusion Based on Edge-Intensities. *10th International Conference on Document Analysis and Recognition*, pp. 136-140, 2009.

[27] T. Mertens, J. Kautz, F. V. Reeth. Exposure Fusion: A simple and practical alternative to high dynamic range photography. *Computer Graphics Forum,* vol. 28(1), pp. 161-171, 2007.

[28] L. Yang, B. Guo, W. Ni. Multi-focus Image Fusion Algorithm Based on Contourlet Decomposition and Region Statistics. *Fourth International Conference on Image and Graphics*, pp. 707-712, 2007.

[29] S. Li, B. Yang. Multifocus image fusion using region segmentation and spatial frequency, *Image and Vision Computing*, vol. 26(7), pp. 971-979, 2008.

[30] L. Ding, C. Han. Multi-focus Image Fusion Using Wavelet Based Contourlet Transform and Region, *International Conference on Information Management and Engineering*, pp. 90-93, 2009.

[31] L. Tang, Z. Zhao. The Wavelet-based Contourlet Transform for Image Fusion, *Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing*, vol. 2, pp. 59-64, 2007.

[32] M. Choi, R.Y. Kim, M.G. Kim. The curvelet transform for image fusion, *20th Congress of the Internet, Society for Photogrammetry and Remote Sensors*, pp. 59-64, 2004.

[33] F. Sroubek, S. Gabarda, R. Redondo, S. Fischer, S. Crist´obal. Multi-focus Fusion with Oriented Windows, *Proceedings of SPIE, Bioengineered and Bioinspired Systems II*, vol. 5839, pp. 264–273, 2005

[34] T. Zaveri, M. Zaveri. A novel two step region based multi focus image fusion method, *International journal of computer and electrical engineering*, vol. 2(1), 2010.

[35] G. Ward. Fast, Robust Image Registration for Compositing High Dynamic Range Photographs from Handheld Exposures, J*ournal of Graphic Tools*, vol. 8, pp. 17-30, 2003.

[36] T. Grosch. Fast and Robust High Dynamic Range Image Generation with Camera and Object Movement, *Proceedings of Vision, Modelling and Visualization*, pp. 277-284, 2006.

[37] A. Tomaszewska, R. Mantiuk. Image registration for multi-exposure high dynamic range image acquisition, *The 15th International Conf. in Central*

*Europe on Computer Graphics, Visualization and Computer Vision,* pp. 49-56, 2007.

[38] M. D. Grossberg, S. K. Nayar. Modelling the Space of Camera Response Functions, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26(10), pp.1272-1282, 2004.

[39] A. Myronenko, X. Song, M. Carreira-Perpinan. Non-rigid point set registration: Coherent point drift, *Proceedings of Advances in Neural Information Processing Systems*, pp. 1009-1016, 2006.

[40] X. Clady, P. Negri, M. Milgram, R. Poulenard. Multi-class Vehicle Type Recognition *System, Proceedings of the workshop on Artificial Neural Networks in Pattern Recognition,* pp. 228-239, 2008.

[41] F. M. Kazemi, S. Samadi, H. R. Poorreza, M. Akbarzadeh. Vehicle Recognition Based on Fourier, Wavelet and Curvelet Transforms - a Comparative Study. *Fourth International Conference on Information Technology,* pp. 939-940, 2007.

[42] F. M. Kazemi, S. Samadi, H. R. Poorreza, M. Akbarzadeh. Vehicle Recognition Using Curvelet Transform and SVM. *Fourth International Conference on Information Technology,* pp. 516-521, 2007.

[43] S. Rahati, R. Moravejian, E. Mohamad, F. Mohamad. Vehicle Recognition Using Contourlet Transform and SVM, *Fifth International Conference on Information Technology New Generations,* pp. 894-898, 2008.

[44] I. Zafar, E. A. Edirisinghe, S. Acar. Localized contourlet features in vehicle make and model recognition. *In Proceedings SPIE, Electronic Imaging*. Vol. 7251(5), 2009.

[45] M. M. Arzani, M. Jamzad. Car type recognition in highways based on wavelet and contourlet feature extraction. *International Conference on Signal and Image Processing*, pp. 353-356, 2010.

[46] S. M. Sarfraz, A. Saeed, H.M. Khan, Z. Riaz. Bayesian prior models for vehicle make and model recognition. *In*: *Proceedings of the 7th International Conference on Frontiers of Information Technology*, 2009.

[47] M. S. Sarfraz, M. H. Khan. A Probabilistic Framework for Patch based Vehicle Type Recognition. *In VISAPP* SciTePress, pp. 358-363, 2011.

[48] A. Psyllos, C. N. Anagnostopoulos, E. Kayafas. Vehicle model recognition from frontal view image measurements. *Computer Standards & Interfaces*. Vol. 33(2), pp. 142-151, 2011.

[49] B. Daya, A.H. Akoum, P. Chauvet. Neural Network Approach for the Identification System of the Type of Vehicle. *In 2010 International Conference on Computational Intelligence and Communication Networks*, pp. 162-166, 2010.

[50] J. Prokaj, G. Medioni. 3-D model based vehicle recognition. *In*: *Workshop on Applications of Computer Vision*, pp. 1-7, 2009.

[51] A. Dempster, N. Laird, D. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society* vol. 39(1), pp. 1–38, 1977.

[52] R. A. Redner, H. F. Walker. Mixture densities, maximum likelihood and the EM algorithm, *SIAM Review*, vol. 26(2), pp. 195-239, 1984.

[53] A. L. Yuille, N. M. Grzywacz. The motion coherence theory, *International Journal of Computer Vision*, vol. 3, pp. 344 - 353, 1988.

[54] A. Myronenko, X. Song. Point-Set Registration: Coherent Point Drift", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 32(12), pp. 2262-2275, 2010.

[55] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints, *International Journal of Computer Vision*, vol. 60(2), pp. 91-110, 2004.

[56] J. Sivic, F. Schaffalitzky, A. Zisserman. Efficient Object Retrieval from videos, *Proceedings of the 12th European signal Processing Conference*, pp. 1737–1740, 2004.

[57] M. N. Do, M. Vetterli. The Contourlet Transform: An Efficient Directional Multi-resolution Image Representation, *IEEE Transactions Image on Processing*, vol. 14(12), pp. 2091-2106, 2005.

[58] M. N. Do, M. Vetterli. Pyramidal directional filter banks and Curvelets, *IEEE Int. Conference on Image Processing*, vol. 3, pp. 158-161, 2001.

[59] P. J. Burt, E. H. Adelson. The Laplacian pyramid as a compact image code, *IEEE Trans. Commun.*, vol. 31(4), pp. 532–540, 1983.

[60] M. A. Fischler, R. C. Bolles. "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography". *Comm. of the ACM*, vol. 24(6), pp. 381–395, 1981.

[61] Y. Freund, R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, vol. 55(1), pp. 119-139, 1997.

[62] M. Sharifi, M. Fathy, M. T. Mahmoudi. A Classified and Comparative Study of Edge Detection Algorithms. *In*: *International Conference on*

*Information Technology: Coding and Computing, 2002. Proceedings*, pp. 117-120, 2002.

[63] P. Viola, M. Jones. Rapid Object Detection using a boosted Cascade of Simple Features, *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition,* vol. 1, pp. 511-518, 2001.

[64] P. Viola, M. Jones. Robust real-time face detection, *International Journal of Computer Vision*, vol. 57(2), pp.137-154, 2004.

[65] S. M. Sarfraz, A. Saeed, H. M. Khan, Z. Riaz. Bayesian prior models for vehicle make and model recognition. *In*: *Proceedings of the 7th International Conference on Frontiers of Information Technology*. 2009.

[66] M. C. Morrone, R.A. Owen. Feature detection from local energy. *Pattern Recognition Letters*. Vol. 6, pp. 303-313, 1987.

[67] A. Yilmaz, O. Javed, M. Shah. Object tracking: A survey. *ACM Computing Surveys*. Vol. 38(4), pp. 1-45, 2006.

[68] R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*. Vol. 82, pp. 34-45, 1960.

[69] T. Perala, R. Piche. Robust Extended Kalman Filtering in Hybrid Positioning Applications. *4th Workshop on Positioning, Navigation and Communication*, Hannover, pp. 55-63, 2007.

[70] OpenCVWiki. 2009. FullOpenCVWiki. [Online] (Updated 10 January 2013) Available at: http://opencv.willowgarage.com/wiki/FullOpenCVWiki [Accessed 03 June 2013].

[71] Andriy Myronenko. Coherent Point Drift (CPD) [Online] (Updated 2010) Available at: https://sites.google.com/site/myronenko/research/cpd/ [Accessed 03 June 2013].

[72] S. Saravi, I. Zafar, E.A. Edirisinghe. Real-time speaker identification for video conferencing. *Real-Time Image and Video Processing 2010, Proceedings of SPIE,* vol. 7724, 2010.

[73] S. Saravi, E. A. Edirisinghe. Contourlet based Multi-exposure Image Fusion with Compensation for Multi-dimensional Camera Shake. *VISAPP* SciTePress, pp. 182-185, 2012.

[74] Lluis-Gomez, S. Saravi, E. A. Edirisinghe. Subjectively optimised multi-exposure and multi-focus image fusion with compensation for camera shake, *Proceedings of SPIE in Optics, Photonics, and Digital Technologies for Multimedia Applications II,* Vol. 8436, 2012.

[75] Wikipedia. 2013. [Online] (Updated 31 October 2012) Available at: http://en.wikipedia.org/wiki/Chroma_subsampling [Accessed 25 January 2013].

[76] Wikipedia. 2013. [Online] (Updated 02 June 2013) Available at: http://en.wikipedia.org/wiki/Mixture_model [Accessed 03 June 2013].

[77] M. S. Sarfraz, et al., Real-time automatic license plate recognition for CCTV forensic applications, *Journal of Real-Time Image Processing*, vol. 69(9), pp. 1-11, 2011.

[78] C. C. Chang, C. J. Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*. Vol. 2(3), 2011.

Available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm [Accessed 03 June 2013].

[79] N. Dalal, B. Triggs. Histograms of Oriented Gradients for Human Detection. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 886-893, 2005.

[80] O. Ludwig. (2011). HOG descriptor for Matlab. [online] (Updated 17 March 2011). Available at: http://www.mathworks.com/matlabcentral/fileexchange/28689-hog-descriptor-for-matlab [Accessed 03 June 2013].

[81] O. Ludwig, D. Delgado, V. Goncalves, U. Nunes. Trainable classifier-fusion schemes: An application to pedestrian detection. *12th International IEEE Conference on Intelligent Transportation Systems*, pp. 1-6, 2009.

[82] Wikipedia. 2013. [Online] (Updated 30 May 2013) Available at: http://en.wikipedia.org/wiki/RANSAC [Accessed 03 June 2013]

[83] Bar-Shalom, Y.; Daum, F.; Huang, J., "The probabilistic data association filter," *Control Systems, IEEE*, vol. 29(6), pp. 82-100, 2009.

[84] Wikipedia. 2013. [Online] (Updated 05 April 2013) Available at: http://docs.opencv.org/modules/video/doc/motion_analysis_and_object_tracking.html [Accessed 03 June 2013].

[85] Photos taken from: http://www.greenmotor.co.uk/2012/05/. [Accessed 03 June 2013].

[86] Wikipedia. 2013. [Online] (Updated 01 June 2013) Available at: http://en.wikipedia.org/wiki/Support_vector_machine [Accessed 03 June 2013].

# Appendix A

## A.1 Publications

The work presented in this thesis has resulted in a number of papers. The list of accepted papers with future publications is listed below:

- S.Saravi et al., 2010. Real-time speaker identification for video conferencing., Real-Time Image and Video Processing 2010, Proceedings of SPIE, 7724, 77240D, 10pp.
- S.Saravi., E.A. Edirisinghe., 2012. Contourlet based Multi-exposure Image Fusion with Compensation for Multi-dimensional Camera Shake., VISAPP (1), 182-185, SciTePress, isbn: 978-989-8565-03-7.
- Lluis-Gomez, S. Saravi, E. A. Edirisinghe, "Subjectively optimised multi-exposure and multi-focus image fusion with compensation for camera shake", in Optics, Photonics, and Digital Technologies for Multimedia Applications II, Proceedings of SPIE Vol. 8436 (SPIE, Bellingham, WA 2012), 84360Q.

## A.2 Publication(s) Under Review/Accepted

- S. Saravi, Lluis-Gomez, E. A. Edirisinghe, "multi-exposure and multi-focus image fusion with camera shake compensation", International Society for Optics and Photonics Journal (SPIE) – Optical Engineering on 14/01/2013 [Accepted]
- S. Saravi, E. A. Edirisinghe, "Vehicle Make and Model Recognition in CCTV Footage", 18th International Conference on Digital Signal Processing 2013, Greece, on 01/02/2013. [Accepted]