# Contents

*Chapter I Introduction* ................................................................................................................................. 1

I.1 Thesis Motivation ........................................................................................................................................ 1

I.2 Outline ......................................................................................................................................................... 4

I.3 Contributions Summary .............................................................................................................................. 6
   I.3.1 Contrast enhancement (photoreceptor) ............................................................................................. 6
   I.3.2 Spatial-temporal pre-compression video enhancement (LMC) ....................................................... 6
   I.3.3 Region of Interest Coding and EMD ............................................................................................... 7
   I.3.4 Face Recognition Enhancement .................................................................................................... 7

*Chapter II Conventional Surveillance Video Technology* ............................................................................... 8

II.1.1 Interdisciplinary research ..................................................................................................................... 8

II.1.2 Research Approach .................................................................................................................................. 8

II.2 General challenges in Surveillance Video .................................................................................................. 10
   II.2.1 Surveillance Video Quality ............................................................................................................... 11
   II.2.2 Complex lighting conditions ............................................................................................................. 12
   II.2.3 Surveillance Cameras Dynamic Range ............................................................................................ 13
   II.2.4 Position of surveillance cameras ..................................................................................................... 14

II.3 Motion in Surveillance Video ..................................................................................................................... 15
   II.3.1 Conventional Motion Detection ....................................................................................................... 16
   II.3.2 Conventional Motion Segmentation .................................................................................................. 17
   II.3.3 Psycho-visual Motion Detection techniques ..................................................................................... 18

II.4 Video Compression ...................................................................................................................................... 18
   II.4.1 Introduction .......................................................................................................................................... 18
   II.4.2 Quality of decompressed Video ........................................................................................................ 19

II.5 General Challenges in Forensic video analysis ......................................................................................... 22
   II.5.1 Surveillance Video quality in forensic analysis ............................................................................... 22
   II.5.2 Video compression and forensic analysis ......................................................................................... 24
   II.5.3 Conventional techniques in forensic analysis ................................................................................... 24

II.6 Summary ..................................................................................................................................................... 25

*Chapter III Biological Vision system* .......................................................................................................... 27

III.1 Human Vision ............................................................................................................................................ 28
| Chapter V LMC and Non-Linear Spatio-temporal Enhancement | 70 |
| V.1 Problem statement | 70 |
| V.1.1 Challenges for surveillance applications | 72 |
| V.1.2 Conventional techniques | 73 |
| V.2 LMC Model Introduction | 75 |
| V.2.2 Post-Compression or Pre-Compression? | 77 |
| V.3 Experiment Description | 79 |

| Chapter IV Photoreceptor and Complex Lighting | 40 |
| IV.1 Problem statement | 40 |
| IV.2 Lighting conditions and Conventional video enhancement | 40 |
| IV.2.1 Histogram equalization | 41 |
| IV.2.2 Tone mapping | 41 |
| IV.3 Biologically Inspired contrast enhancement | 42 |
| IV.4 Brief discussion of time complexity | 44 |
| IV.5 Photoreceptor Model Description | 46 |
| IV.5.1 Gain Control (Stage 1) | 49 |
| IV.5.2 Divisive feedback temporal filtering (Stage 2 and Stage 3) | 49 |
| IV.5.3 Naka-Rushton transformation (Stage 4) | 51 |
| IV.6 Temporal Filtering (Step Response) | 52 |
| IV.7 Photoreceptor Implementation | 53 |
| IV.8 Card Recognition Experiment | 54 |
| IV.8.1 Automatic object recognition | 54 |
| IV.8.2 Shape classification algorithm | 56 |
| IV.8.3 Polar system implementation | 59 |
| IV.8.4 Card segmentation process in real world | 61 |
| IV.8.5 Spatial frequency Analysis | 62 |
| IV.8.6 Statistical Analysis | 65 |

| Chapter III Research Workflow | 38 |
| III.2 Insect Vision systems | 32 |
| III.2.1 Photoreceptor (phototransduction) Features | 35 |
| III.2.2 Laminar Monopolar Cells (LMCs) Features | 36 |
| III.2.3 Elementary Motion Detectors (EMDs) features | 37 |
| III.3 Research Workflow | 38 |

| IV.8 Card Recognition Experiment | 54 |
| IV.8.1 Automatic object recognition | 54 |
| IV.8.2 Shape classification algorithm | 56 |
| IV.8.3 Polar system implementation | 59 |
| IV.8.4 Card segmentation process in real world | 61 |
| IV.8.5 Spatial frequency Analysis | 62 |
| IV.8.6 Statistical Analysis | 65 |
V.3.1 Fourier Descriptor: ................................................................. 82
V.3.2 Video Compression Implementation ........................................ 84

V.4 Statistical analysis ........................................................................ 85
V.4.1 Comparison with Conventional Methods ...................................... 89

Chapter VI EMD and Region of Interest Coding ................................. 91
VI.1 Problem statement ....................................................................... 91
VI.1.1 Challenges for surveillance systems .......................................... 91
VI.1.2 Region of Interest Coding ....................................................... 92
VI.1.3 Conventional enhancement techniques ...................................... 93

VI.2 Elementary Motion Detection (EMD) Model Description ............... 95

VI.3 EMD Implementation .................................................................. 96
VI.3.1 Experiment Description ............................................................ 96

Chapter VII Integration – Early study on facial recognition .................... 107
VII.1 Introduction ................................................................................ 107

VII.2 Face Recognition and Compressed Video ..................................... 110
VII.3 Methodology .............................................................................. 112
VII.3.1 Introduction to PCA Algorithm .............................................. 115

VII.4 Results ...................................................................................... 118

Chapter VIII Conclusion and Future Work ......................................... 121
VIII.1 Discussion of Challenges ........................................................... 121
VIII.1.1 Surveillance video compression .............................................. 121
VIII.1.2 Biological vision system ......................................................... 121
VIII.1.3 Insect vision system and its components ................................... 121

VIII.2 Future Direction ....................................................................... 122

Appendix A ....................................................................................... 124
A.1 Filtering in Spatial Domain ............................................................ 124
A.1.1 Canny Edge detector ................................................................. 124
A.1.2 Symmetric Gaussian low pass filter ........................................... 125

A.2 Template Matching ...................................................................... 126

A.3 Thresholding Techniques ............................................................ 127

A.4 Affine Transformation .................................................................... 129
A.4.1 Translation: ........................................................................................................... 130
A.4.2 Rotation: ............................................................................................................... 130
A.4.3 Scaling: .................................................................................................................. 130
A.4.4 Reflection: ............................................................................................................ 130

Appendix B Matlab Codes ............................................................................................. 132

B.1 Bio-Inspired Pre-Processing Algorithm .................................................................. 132
  B.1.1 Main code ............................................................................................................ 132
  B.1.2 Photoreceptor ..................................................................................................... 134
  B.1.3 LMC .................................................................................................................... 135
  B.1.4 EMD .................................................................................................................... 136
Abstract

In this thesis, the application of biomimetic vision models is proposed and evaluated in the field of surveillance video enhancement. It is argued that conventional video compression and representation, even that which is used in surveillance applications, is optimised for entertainment purposes and is demonstrably compromised when it comes to retention of details of relevance to recognition of surveillance-relevant objects such as faces and car licence plates.

Four sets of investigations with experimental results are presented. These are the application of three stages of biomimetic modelling of the blowfly eye and psychovisual system:

1. The Photoreceptor Model as a non-linear temporal enhancement method. It is demonstrated that the contrast enhancement introduced by this process improves object recognition under real-world lighting conditions, with specific application to the recognition of shapes (i.e. playing card suits in our experiments).

2. The Laminar Monopolar Cell (LMC) model as a non-linear spatio-temporal information compression stage. This stage retains, in particular, the details of moving objects in the field of view. The application of this stage to car licence plate alpha-numeral characters is demonstrated as a pre-processing stage before conventional MPEG-like video compression is applied. It is shown that under low to moderate levels of video compression and under realistic lighting conditions, that distinguishing features between similar characters are retained, hence improving the performance of subsequent character recognition.

3. Elementary Motion Detection (EMD) as a subsequent biomimetic stage which measures velocity in the field of view. The EMD is applied as a detector of moving objects in the field of view, which are subsequently investigated as a Region of Interest in surveillance applications. It is demonstrated under complex lighting conditions that car licence plate details can be retained at high compression rates using this approach, especially when combined with LMC enhancement, compared with conventional approaches with the same data bandwidth constraints.

4. The LMC and EMD models are also considered in a preliminary study of facial feature enhancement and recognition. It is demonstrated that facial features are retained at lower data rates than conventional signal processing approaches would support.

Results are compared with conventional signal-processing based enhancement approaches, and computational complexity is also considered. It is argued and demonstrated that the biomimetic approach is not only effective in improving recognition rates through the retention of structural details in enhanced video sequences, but that the enhancement is of relatively low computational complexity, and is highly suited to contemporary parallel graphics processing.
Statement of Originality

This work contains no material that has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and believes, contains no material previously published or written by another person, except where due reference has been made in the text.

I give consent to this copy of my thesis being available in the University Library.

The author acknowledge that copyright of published works contained within this thesis (as listed under publications) resides with the copyright holder/s of those works.

Signed

Date
Acknowledgments

Foremost, I would like to express my sincere gratitude to my supervisors Dr Matthew Sorell and Dr Russell Brinkworth for the continuous support of my PhD study, for their patience, motivation, enthusiasm, and immense knowledge. I have been extremely lucky to have supervisors who cared so much about my work, and who responded to my questions and queries so promptly. Besides my supervisor, I would like to thank Dr. Brian Ng for encouragement and insightful comments.

I would like to thank the School of Electrical and Electronic Engineering for all the resources that have been made available to aid me in my research. Completing this work would have been all the more difficult were it not for the support and friendship provided by people working in this School. I am indebted to them for their help. I would also like to thank all the members of staff at University of Adelaide.

I would like to thank my family for all their love and encouragement especially my parents and my brother (Kamal Poursoltanmohammadi, Farnaz Niroumand Zandi and Shayan Poursoltan) who supported me spiritually in my life. I was continually amazed by their patience throughout all of the ups and downs of my research. Finally, I would like to thank the anonymous reviewers, for taking their valuable time to review this manuscript. Their constructive comments have been of tremendous value.
Publications


List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>II-1</td>
<td>This research areas diagram in the literature (1) and how they could be addressed (2)</td>
<td>10</td>
</tr>
<tr>
<td>II-2</td>
<td>Video encoder impacts on license plate recognition</td>
<td>21</td>
</tr>
<tr>
<td>II-3</td>
<td>A motion-blurred video frame and de-blurring techniques</td>
<td>25</td>
</tr>
<tr>
<td>III-1</td>
<td>Simplified cross-sectional diagram of the human eye</td>
<td>27</td>
</tr>
<tr>
<td>III-2</td>
<td>Block diagram of conventional video compression algorithms</td>
<td>30</td>
</tr>
<tr>
<td>III-3</td>
<td>Video technology development</td>
<td>31</td>
</tr>
<tr>
<td>III-4</td>
<td>Structure of building block of the fly’s retina</td>
<td>33</td>
</tr>
<tr>
<td>III-5</td>
<td>Overview of insect vision stages studied in this research</td>
<td>34</td>
</tr>
<tr>
<td>III-6</td>
<td>Temporal responses of photoreceptors and LMCs</td>
<td>37</td>
</tr>
<tr>
<td>IV-1</td>
<td>Conventional methods vs photoreceptor</td>
<td>43</td>
</tr>
<tr>
<td>IV-2</td>
<td>Photoreceptor Model diagram</td>
<td>47</td>
</tr>
<tr>
<td>IV-3</td>
<td>Photoreceptor experiment</td>
<td>48</td>
</tr>
<tr>
<td>IV-4</td>
<td>This figure shows the processing pathway for a simplified photoreceptor model</td>
<td>49</td>
</tr>
<tr>
<td>IV-5</td>
<td>Coefficients impacts on Photoreceptor Model Stage Step response</td>
<td>50</td>
</tr>
<tr>
<td>IV-6</td>
<td>Stage 1 response in darker and lighter areas</td>
<td>50</td>
</tr>
<tr>
<td>IV-7</td>
<td>Photoreceptor Model Stage Step response</td>
<td>51</td>
</tr>
<tr>
<td>IV-8</td>
<td>Stage 4 of photoreceptor model</td>
<td>52</td>
</tr>
<tr>
<td>IV-9</td>
<td>Flowchart of employed bio-inspired video enhancement method</td>
<td>53</td>
</tr>
<tr>
<td>IV-10</td>
<td>Three dimensional spatiotemporal structure of video</td>
<td>55</td>
</tr>
<tr>
<td>IV-11</td>
<td>Playing card suits</td>
<td>57</td>
</tr>
<tr>
<td>IV-12</td>
<td>Card Suit Classifier Algorithm Flowchart</td>
<td>58</td>
</tr>
<tr>
<td>IV-13</td>
<td>Generated synthetic video frames with different affine transformation</td>
<td>59</td>
</tr>
<tr>
<td>IV-14</td>
<td>Boundary signature in Polar Coordinate</td>
<td>61</td>
</tr>
<tr>
<td>IV-15</td>
<td>Card Segmentation Algorithm</td>
<td>63</td>
</tr>
<tr>
<td>IV-16</td>
<td>Forward one-level 2D DWT via sub-band coding scheme</td>
<td>64</td>
</tr>
<tr>
<td>IV-17</td>
<td>Multiresolution structure of wavelet decomposition of an image</td>
<td>65</td>
</tr>
<tr>
<td>IV-18</td>
<td>Wavelet Transform (WT) frequency analysis</td>
<td>66</td>
</tr>
<tr>
<td>IV-19</td>
<td>The confidence interval of each method is depicted for better view of their performance</td>
<td>68</td>
</tr>
<tr>
<td>IV-20</td>
<td>Showing frames enhanced by photoreceptor model(above) versus unenhanced (original) frames</td>
<td>69</td>
</tr>
<tr>
<td>V-1</td>
<td>Confusion in number plate recognition under JPEG compression</td>
<td>71</td>
</tr>
<tr>
<td>V-2</td>
<td>Typical automatic license plate recognition system</td>
<td>74</td>
</tr>
<tr>
<td>V-3</td>
<td>Algorithm and schematic of implemented LMC Model (Note: High pass filter in step 3 is temporal)</td>
<td>76</td>
</tr>
<tr>
<td>V-4</td>
<td>Using LMC for Post-Compression and Pre-Compression</td>
<td>78</td>
</tr>
<tr>
<td>V-5</td>
<td>Using LMC model as post-processing algorithm (character: ‘G’), quantization scale =3)</td>
<td>78</td>
</tr>
<tr>
<td>V-6</td>
<td>Character Fourier Descriptors</td>
<td>79</td>
</tr>
</tbody>
</table>

XIII
FIGURE V-7 FE-SCHRIFT font that were used in German Number Plates ................................................................. 80
FIGURE V-8 SYNTHETICALLY DEVELOPED VIDEO (fr=frame number) ................................................................. 82
FIGURE V-9 FOURIER DESCRIPTOR WITH DIFFERENT NUMBER OF M (frequency components) .................. 83
FIGURE V-10 FLOWCHART OF DCT BASED INTRA-FRAME CODING IN MPEG-2 ........................................... 84
FIGURE V-11 CHARACTERS AUTOCORRELATION COMPARISON ........................................................................... 87
FIGURE V-12 SEPARATION ABILITY COMPARISON IN ORDER TO SEE THE BIOLOGICAL ENHANCEMENT IMPACT ............................................................... 88
FIGURE V-13 - PERCENTAGE OF THE TIME POINTS SHOWING IMPROVEMENT FOR EACH CHARACTER .......... 89
FIGURE VI-1 STRUCTURE OF A ELEMENTARY MOTION DETECTOR .................................................................. 96
FIGURE VI-2 LABORATORY SET UP ......................................................................................................................... 97
FIGURE VI-3 IMPLEMENTED EMDs (COMBINED WITH LMCs) ALGORITHM ...................................................... 98
FIGURE VI-4 PRE-PROCESSED AND COMPRESSED FRAMES COMPARISONS .................................................. 99
FIGURE VI-5 PSNR COMPARISON BETWEEN THE RECONSTRUCTED VIDEO. ....................................................... 100
FIGURE VI-6 SEGMENTED OBJECT PSNR COMPARISON OF ENCODED VIDEO (MPEG-2) WITH AND WITHOUT PRE-PROCESSING ........................................... 101
FIGURE VI-7 PSNR AND AUTOCORRELATION COMPARISON FOR DIFFERENT DATA RATE VIDEOS ................. 102
FIGURE VI-8 AVERAGE AUTOCORRELATION COMPARISON FOR DIFFERENT CHARACTERS IN ............................................................... 103
FIGURE VI-9 FLOWCHART OF COMPARISON METHOD (PROPOSED VS. CONVENTIONAL) .......................... 104
FIGURE VII-1 TYPICAL EXAMPLE OF AVAILABLE SURVEILLANCE VIDEO ....................................................... 108
FIGURE VII-2 FACE RECOGNITION SYSTEM DIAGRAM ......................................................................................... 111
FIGURE VII-3 THE VIDEO FRAMES CAPTURED FOR THIS WORK ........................................................................ 114
FIGURE VII-4 TRAINING SET OF THIS EXPERIMENT WHICH IS OBTAINED FROM Libor Spacek DATABASE [248] .................. 115
FIGURE VII-5 EIGENFACES THAT ARE DERIVED FROM TRAINING SET .......................................................... 116
FIGURE VII-6 SCHEMATIC OF PCA FACE RECOGNITION ALGORITHM .......................................................... 118
FIGURE VII-7 PCA FACE RECOGNITION ACCURACY VS. COMPRESSION RATIO .................................................. 119
FIGURE VII-8 MAXIMUM EUCLIDEAN DISTANCE ................................................................................................. 120
FIGURE A-0-1 DISCRETE APPROXIMATION TO GAUSSIAN FUNCTION WITH $\sigma = 1.0$ ........................................ 126
FIGURE A-0-2 DIFFERENT TYPES OF AFFINE TRANSFORMATIONS .................................................................... 131

XIV
List of Tables

TABLE III-1 SUMMARY OF THE DIFFERENCES BETWEEN THE RODS AND CONES [120] ................................................................. 29
TABLE IV-1 TIME COMPLEXITY OF THE CONVENTIONAL AND BIOLOGICAL METHODS ................................................................. 45
TABLE IV-2 RECOGNITION RATE COMPARISONS BETWEEN ORIGINAL AND ENHANCE VIDEO FRAME .................................................. 67
TABLE V-1 ALPHA-NUMERAL CHARACTERS CORRELATION MATRIX (WITHOUT COMPRESSION) .............................................................. 81
TABLE V-2 RECOGNITION RATE FOR CHARACTER RECOGNITION ........................................................................................................... 90
TABLE VI-1 OVERALL PERFORMANCE COMPARISON (CONVENTIONAL VS. BIOLOGICAL) ................................................................. 105
TABLE VI-2 CHARACTER RECOGNITION RESULTS ............................................................................................................................. 106
Chapter I  INTRODUCTION

1.1 THESIS MOTIVATION

In this thesis, we consider the application of Insect Vision Modelling to the challenge of improving the storage and analysis of video, with particular interest in surveillance. Unlike conventional image processing approaches, insect vision exploits both temporal and spatial features and lends itself to low computational complexity in a highly parallel processing environment. We are motivated by the ability of insects to enhance contrast of moving objects in highly cluttered scenes, and to track the motion of such objects. We therefore investigate whether introducing insect-vision model approaches can improve object contrast, particularly under challenging lighting conditions, by considering the performance of object classification algorithms (such as alpha-numeral and face recognition); and the use of biomimetic motion detection as a means of implementing Region-of-Interest coding. These techniques are also considered in the context of lossy video compression as part of a surveillance system.

A short surveillance video of a person of interest may be released to the media by law enforcement agencies in order to obtain more information from the general public. However, often there is a lack of recognisable features in regions of interest [1]. Analysis of such video, whether in a public call for information or through automated or expert processes, is therefore of limited use. In many cases, such objects (e.g. number plates, faces or many of the details of clothing of a suspect etc.) are not recognisable. Poor quality videos have a strong chance of being excluded in courts and a case may be lost if this is the only key evidence [1]. Furthermore, the artefacts introduced by sophisticated processing in video surveillance systems, such as compression, complicate the scenario by discarding required information [2] and such unwanted artefacts can threaten the usefulness and fidelity of such evidence [3]. It has been demonstrated that lossy compression alone can be the main source of information alteration [4]. There has not been a great effort to design specialised video compression for surveillance applications and there is a significant technology gap in surveillance video coding [5].

Conventional video compression introduces limitations by which the video quality can be questioned in law courts [6]. In fact, most contemporary video encoders and video quality metrics were originally designed for television and telecommunication applications. Using the same approaches in surveillance systems raises many issues of fidelity and trust.

Commonly used video encoding algorithms are mostly designed to cater for biological systems (that is, human vision), not to mimic biological processes (whether human or other simpler vision systems). The goal in this research work is to investigate how we can utilize biomimetic vision systems in conventional surveillance video encoding systems to resolve abovementioned issues and
CHAPTER I INTRODUCTION

hence improve the resulting surveillance video quality. The challenge is that conventional video quality metrics are concerned with entertainment purposes, which have significantly different requirements to surveillance applications. Efficacy in surveillance video analysis is considered in the context of the identification of objects and the detection of events [7]. In contrast, entertainment-based video is regarded as high quality if the non-expert viewer has a pleasant experience watching the video [8].

This work can be classified as interdisciplinary research as it needs to interconnect different yet relevant major fields. The bridge between disciplines, while narrow, can be built deeply on the foundations of the traditionally distinct contributing fields. In fact, insights into complex problems often require contributions from a range of disciplines. The three major fields which have been linked in this research are:

I. Surveillance video analysis (related to expert interpretation of evidence)
II. Video compression (related to implementation of video compression technology)
III. Biological vision systems (related to the study and mimicry of biological systems)

Conventionally these fields have been largely isolated from each other, drawing only the most obvious connections in the simplest ways. It was therefore necessary to develop a sufficiently deep understanding of each of these three fields, their connections and the potential gaps which could be addressed. It has been identified that such relevant topics have not been as strongly interconnected as might be expected in the literature. In order to link these fields it was necessary to narrow down the research to a significant application area which has not been thoroughly addressed in the literature. Hence, this work is limited to scenarios in which forensic video analysis often encounters significant challenges.

For example, consider the challenges which arise under complex lighting conditions such as deep shadow or bright backgrounds. Contemporary surveillance cameras are generally low cost, have poor dynamic range, and rely on global settings to enhance contrast. They are generally not very efficient under low lighting conditions, resulting in either loss of spatial detail or, if long integration times are used, temporal smearing.

Biological vision systems have a range of techniques for adapting to a variety of lighting conditions [9] and this feature enables animals, including humans, to see objects and their details, even in very dark, extremely bright and high contrast (combination of bright and dark) environments. This biological vision system feature has inspired us to study and utilise it in conventional surveillance systems. In the work presented here, the photoreceptor model derived from analysis of flies (e.g. hoverflies) [10] is considered as a highly parallel, non-linear temporal compression algorithm for improving contrast enhancement. In Chapter IV this approach is implemented as a pre-processing stage in the context of playing card suit recognition under noisy and complex lighting conditions with video captured under realistic scenarios. It is demonstrated that even with non-optimised parameter
settings, classification performance is substantially improved. In contrast, the conventional image-processing approaches of tone-mapping and histogram equalisation do not result in a statistically significant improvement in classifier performance.

This approach is taken further with a second biomimetic stage known as Laminar Monopolar Cells [11], which introduces spatio-temporal processing to suppress noise and further enhance contrast features of moving objects in the field of view. In the process, still background features are suppressed as a form of data compression. In Chapter V the LMC processing stage is considered as a form of enhancement of alpha-numerical characters as would be found on a vehicle licence plate. It is demonstrated that characters which are most susceptible to confusion (such as H versus N), especially when lossy compression artefacts are introduced, become more readily distinguishable, as demonstrated by an increase in correlation against a character template, and a decrease in cross-correlation between the similar (but different) characters. This enhancement works for low to moderate levels of lossy compression. For highly compressed video, however, it is shown that there is no improvement achieved.

A review of the literature also suggests that biological models, while forming the basis of video compression, are no longer aligned with developments of video compression technology [12]. Conventional video compression algorithms treat the entire video frame uniformly. The concept of region-of-interest (ROI) coding however is introduced to treat areas in a frame unequally and therefore facilitate different strategies for encoding each classified area [13]. Region of Interest (ROI) coding requires an effective algorithm to identify which part of a video frame in a surveillance application is of particular interest. Therefore, we investigated how additional features of the insect vision model can be used to enhance and improve ROI coding in conventional video compression algorithms. Semantic content reduction in the region of interest (e.g. face, object textures etc.) not only degrades the perceived quality but also might be misinterpreted by people who are monitoring or investigating a crime scene [14]. This is a sensitive subject as this content variation can even occur in seemingly benign video processing stages. Such errors are difficult to perceive in regular normal-speed viewing, yet still have significant impact on post-hoc human (or computer) recognition ability. In fact, the best approach to resolve this issue is arguably to design a surveillance video system in which the ROI has been minimally impacted by different processing stages [15]. Some loss of information however is acceptable and removing nonessential detail from the original can lead to saving a considerable amount of data. It should be noted that ROI is determined by application. For example, the face is of interest in human tracking or video-conferencing, and license plates are of interest in monitored parking areas [16]. Moving objects and their details are commonly of great importance in surveillance applications and hence moving objects have been chosen as ROI targets in this research work.
CHAPTER I INTRODUCTION

Many surveillance systems use available network bandwidth to transfer captured video to the user. Therefore, encoders with data rate control are widely used in such systems. Such encoders dynamically adjust parameters to achieve a target bitrate. A limited budget of bits is the key factor to determine quantization scales in each frame (or group of frames) [17]. In other words, the encoder’s freedom to allocate more bits has been limited by network constraints. The retention of higher spatial frequency regions in DCT-based video encoding results in more coded bits and compressing such regions can result in more quantization and prediction errors. Low pass filtering (effectively, blurring) of regions in which we have little or no interest can give the encoder more freedom to allocate the remaining bits to the ROI. In other words, a smaller Quantization Scale (QS) will be used for frames in which background high frequency details are suppressed and eventually the number of bits allocated to the ROI will be increased at the expense of a decrease in the quality of the background.

While there has been significant progress in the conventional image processing literature [14, 18] to isolate moving objects from a quasi-stationary background, these techniques are generally computationally intensive and demonstrate poor performance under complex lighting conditions [19]. In contrast, the work presented here uses biomimetic Elementary Motion Detection after the LMC enhancement stage on uncompressed surveillance video to separate moving regions of interest from the background.

For the purpose of illustrating biomimetic techniques in this work, alphanumeric character recognition is set as the main task. Face recognition also receives some preliminary consideration to demonstrate the feasibility of the biomimetic approach to this problem. The research demonstrates through a series of experiments that there is a statistically significant improvement in object recognition rates and related classification metrics compared with conventional contrast enhancement techniques.

It is argued and demonstrated that the biomimetic approach is not only effective in improving recognition rates through the retention of structural details in enhanced video sequences, but that the enhancement is of relatively low computational complexity and is highly suited to contemporary parallel graphics processing.

1.2 OUTLINE

Chapter I (this chapter) explains the research motivation. Chapter II describes the technical background on which the work in this thesis is based. The chapter starts with a short history of video surveillance and video compression and how these have developed. It then outlines common issues in surveillance videos that have not been fully addressed in the literature. Critical video surveillance factors that can be important in information gathering are discussed and video quality from the forensic perspective is introduced. The impact of lighting conditions on surveillance video quality is
also described. Conventional techniques for block-based video compression and motion detection are introduced, and the limitations of these techniques in video surveillance are discussed.

Chapter III introduces biomimetic vision system models with an emphasis on insect vision system elements. The insect vision system model is comprised of different elements of which photoreceptors, Large Monopolar Cells (LMC) and Elementary Motion Detection (EMD) are studied and implemented in this work. Each of these elements has its own unique features that enables it to be applied to relevant examples. This chapter is introductory in nature, with details of the element models and implementation presented in later chapters.

Chapters IV, V and VI are dedicated to demonstrating the applications of the photoreceptor, LMC and EMD models respectively. The main experiments performed in this research work are thoroughly explained in these chapters. These experiments are specifically designed to demonstrate how each biological vision system component can help to improve the forensic video quality. Note that some of these elements can work better when they are combined. Hence, the main element of interest is coupled with earlier elements in Chapter V and VI. Three experiments are described and this leads into the description of new methodologies and how our proposed new algorithms contribute to solve previously stated issues (in Chapter II and III).

In the first set of experiments (Chapter IV) a playing card suit recognition system was developed. The first biomimetic insect vision system component (i.e. photoreceptor) was implemented as a pre-processing stage to the classification system to improve the classification performance. It is demonstrated that such pre-processing results in a statistically-significant improvement in detector performance, even with ad-hoc non-optimised parameter selection.

In the second set of experiments (Chapter V), the Laminar Monopolar Cells (LMCs) model was used as an advanced spatiotemporal video enhancement approach. The impact of LMC enhancement on alpha-numeral character recognition pre- and post- video compression is considered, showing that pre-enhancement by LMC of, for example, licence plates, can substantially improve character recognition rates for low to moderate compression rates.

In Chapter VI the idea of using biologically inspired methods for region of interest coding in surveillance cameras is proposed. The third insect vision system component (i.e. EMDs) was used as a novel pre-compression video enhancement to overcome deteriorating effects that can be caused by video encoders in regions of relatively high motion. Of particular novelty is that the biomimetic model can either be used to pre-enhance video content for compression using black box video compression of an off-the-shelf surveillance system, or can alternatively manipulate compression parameters of a known compression system to improve retention of forensically-relevant details.

Since the proposed bio-inspired pre-processing showed its advantages in pattern recognition applications, a preliminary study was conducted to identify whether face recognition systems could also benefit from the proposed biologically-inspired video compression and initial experiments and
results are included in Chapter VII. Finally, future work and conclusions resulting from this research are considered in Chapter VIII.

I.3 CONTRIBUTIONS SUMMARY

The main contributions of this thesis are demonstrations that biomimetic techniques, based on insect vision models, are effective and efficient mechanisms for enhancement of features in surveillance videos in order to improve content recognition tasks. These contributions are briefly discussed as follows:

I.3.1 Contrast enhancement (photoreceptor)
A novel adaptive contrast enhancement method inspired by photoreceptor has been proposed [20]. In contrast to other conventional contrast enhancement, this method exploits temporal information rather than spatial information. It is demonstrated that this method can improve shape identification systems (i.e. card suit recognition system in our experiment) performance under a wide variety of lighting conditions. Since the photoreceptor model adaptively compresses intensity variation, bright and dark regions receive much the same attention for further processing, efficiently mitigating lighting variations [21]. Our experimental results show that the proposed method can significantly improve the card suit recognition performance, and yet under the same conditions the use of a widely used conventional approach shows only marginal improvement. A comparison of this bio-inspired method to conventional ones also suggests that this method can be less computationally expensive and hence can be suited real time applications. The developed card suit recognition system in our experiment was also novel and its invariance and accuracy in various form of affine transformation is demonstrated in chapter IV.

I.3.2 Spatial-temporal pre-compression video enhancement (LMC)
A novel pre-compression video enhancement method is proposed which is inspired by the properties of the second component of insect vision model (i.e. LMC) [3]. It is discussed and demonstrated that this method can significantly improve compressed video quality in low light conditions to ensure sufficient details for later analysis has been captured. This pre-compression stage effectively enhances changes in space and time, thereby highlighting the object boundary points. Our hypothesis is that more semantic information would be retained at comparable data rates, thereby improving recognition rates. We also demonstrate that integration of our proposed method can lead to recognition accuracy improvement as it will help the system to retain important structural detail of characters, even after lossy compression. The unique feature of this method is it creates a distinctive border between character edges and background. Our experimental results also show the biological method is capable of improving character recognition systems accuracy in comparison with conventional methods, specifically in complex lighting conditions.
1.3.3 Region of Interest Coding and EMD

Novel bio-inspired ROI video coding is proposed in order to help retain vital information in highly compressed video sequences with high spatial frequency details and motion contents.

We demonstrate that the number of bits allocated to the ROI can be effectively increased at the expense of a decrease in the quality of the background. Although a similar idea has been previously proposed in the literature [22], a biomimetic vision model, to our knowledge, has never been employed in ROI segmentation. In this work, integration of LMC and EMD is employed to segment the ROI from background. In fact, a joint system is presented that is comprised of a pre-processing stage (i.e. biologically inspired ROI segmentation) and a conventional encoder. It is demonstrated that the proposed biologically-inspired segmentation method has advantages over existing methods. It was demonstrated that this biomimetic approach offers higher performance in low light conditions and lower computational complexity. Object recognition systems can significantly benefit from such features [23].

1.3.4 Face Recognition Enhancement

The possibility of using the proposed biological ROI segmentation method to improve accuracy of face recognition tasks in surveillance videos is discussed and studied in last chapter (Chapter VII). The impacts of video compression artefacts on face feature were firstly investigated and it was then demonstrated that conventional face recognition system accuracy can be improved by using our proposed biologically-inspired method.
Chapter II  CONVENTIONAL SURVEILLANCE VIDEO TECHNOLOGY

II.1.1 Interdisciplinary research
Insights into complex problems often require contributions from a range of disciplines. The well-known challenge is to bridge not only the knowledge gap, but also the significant cultural differences including nomenclature, approaches to the development of experiments and the testing of hypotheses, and the discipline-specific foundation research which largely defines modern discipline-specific practice.

Language presents the biggest challenge, because language is built on cultural preconceptions which are often not appreciated by practitioners in a particular discipline. It takes time and patience to develop a common framework which retains the necessary depth of understanding from each contributing field. It is very tempting, but futile, to dumb-down the foundations in this case.

The approach in the research presented here is to focus on a relatively narrow applied topic, improving object recognition in video footage by using biologically-inspired enhancement, and so make it clear that the bridge between disciplines, while narrow, also builds deeply on the foundations of the traditionally distinct contributing fields.

In the process, it becomes apparent that not only could novel solutions be put forward to traditionally complex problems, but that this could arise relatively efficiently, by drawing on the approaches developed in one discipline but applied to a problem identified in another.

II.1.2 Research Approach
It has been identified that some of the relevant topics for this research, including video surveillance, biological vision system modelling, video compression and forensic video analysis have not been as strongly interconnected as might be expected in the literature. The purpose of this chapter is to highlight and introduce the diverse technologies associated with video surveillance and the limitations these technologies face in forensic analysis. In the next chapter, biological vision models will be introduced and it will be argued that such models can be applied to address some such limitations.

The experimental research presented in subsequent chapters demonstrates how such biomimetic models can result in statistically significant improvements in the performance of forensic analysis tasks, particularly pattern and object recognition.

The four research areas of interest in this research are:

- Video surveillance, including camera optics and lighting
- Video forensic analysis, including pattern recognition
- Video compression, including region-of-interest coding
- Biological vision systems, including synthesis
CHAPTER II CONVENTIONAL SURVEILLANCE VIDEO TECHNOLOGY

Each of these topics has received intensive attention within their own fields of study, and it is neither feasible nor necessary to provide a complete background in each topic.

On the other hand, the links between these fields have received little, if any, attention in the existing academic literature. For example, video compression research is mostly concerned with algorithms targeting the entertainment media environment. Video compression, albeit based on standards optimised for entertainment, has been an intrinsic part of video surveillance systems for a long time and is still employed in video surveillance technology [24]. However, the impact of this stage on video quality from the forensic analysis perspective has, to the best of our knowledge, never been thoroughly studied in the academic literature.

The modelling of biological vision systems, on the other hand, is an established study in the field of biology. Extensive research has been published of studies seeking to explain biological vision system behaviours [25]. However, the application of this science to practical engineering of improved electronic vision systems of any kind has been limited when it comes to the gap between such science and existing video system engineering, which is based on relatively simple biological characteristics of vision such as resolution, frame rate, and colour representation [26]. It is argued here that more recent advances in biomimetic vision systems can be applied to the ongoing development and improvement of video representation and compression systems.

Video forensic analysis is a term which broadly refers to two areas of specialisation: the examination of the content of videos as evidence, and the evaluation of video quality and its suitability as legal evidence. The latter study is concerned with a wide range of matters including camera placement, lighting conditions, and artefacts potentially introduced by compression. This field includes video footage derived from a wide variety of sources, including surveillance systems, crime scene photography, and that supplied by, or acquired from, members of the public. The relatively recent rise of digital still and video cameras, including integrated cameras in communication devices, has been driven by consumer applications, leaving the study of forensic applications, based on analogue electronic technology and film, at a disrupted disadvantage.

Figure II-1 summarises where the research in this thesis seeks to expand on the relatively small body of literature in the overlaps between video as a surveillance and forensic analysis tool, video compression, and biological vision system modelling. A review of the literature suggests that biological models, while forming the basis of video compression, are no longer aligned with developments of video compression technology. Region of Interest coding, while fitting both surveillance applications and video compression techniques, is not exploiting the overlap to its full potential. Finally, video forensic analysis is currently embedded within the study of video as an evidentiary tool, whereas the broader question, that of the suitability of video as legal evidence, needs to draw more heavily on developments in both video compression and biological modelling.
At the time of writing, there are clear gaps in the existing literature when it comes to these overlapping fields. The research presented here seeks to expand, specifically, the application of biomimetic vision models in video compression, with a specific focus on video surveillance systems. In doing so, it is also argued that it expands the existing literature on video forensic analysis using biomimetic tools, biomimetic applications for region-of-interest coding, and the interdependencies between surveillance applications and video compression even without direct reference to biological models.

The research seeks to demonstrate through a series of experiments the merit of applying biomimetic models in a more sophisticated way to video compression technology developments, using video surveillance and forensic analysis as the vehicle.

II.2 General challenges in Surveillance Video

Surveillance systems are sometimes referred to as eyes on the street. At their best they can support or enhance traditional investigation. In the hands of trained and experienced investigators, their images can provide vital intelligence. Good quality video – whatever that means – can help investigators and
officers see incidents and people of interest. They can assist in witness interviews and corroboration. On the other hand, if a surveillance system does not deliver clear images of items of interest (e.g. car licence plates and similar objects), then the captured surveillance video has diminished power as legal evidence in court.

Surveillance systems have been closely aligned with commercial development of television cameras and recording equipment. In the United States, the practice moved from dedicated strategic locations to wider use around 1965 [27]. Video tape (and later, video cassettes) was commonly used as the primary storage medium, developing the idea of preservation of evidence on analogue media. In Britain, video surveillance systems began to be installed in major underground train stations in the early 1970s. The idea of traffic flow monitoring by fixed cameras also developed at around the same time [28].

Surveillance systems established themselves more quickly in Britain than the United States. However, the US eventually made up for lost time by implementing more advanced systems in public areas [27]. Analogue systems which were widely used around the world had significant limitations [29]. Other than having very low quality video and low frame rates, users had to change the tapes daily. Many of these problems were resolved in the 1990s by using digital multiplexing. Newer systems offered simultaneous recording on several cameras as well as other features including motion-only recording, which could save considerable tape space. Digital video technologies introduced a new era in the video surveillance industry by improving video quality, employing high performance video encoders and adding more features to the system. These systems allowed users to record higher quality surveillance videos on hard drives. In addition they allowed manipulation of images which made it more useful for different range of applications (such as quality enhancement, pattern recognition systems, etc.) [30].

Consumers and security professionals are now able to view and monitor activities from anywhere in the world by connecting inexpensive cameras to the Internet [31]. Every month, it seems a smaller, more powerful (and possibly cheaper) surveillance system is designed and released out into the market. However, these advanced systems will bring new challenges to security and legal systems that need to be examined carefully which are outside the scope of this study.

Generally, a principle goal of most surveillance cameras is to reduce crime [32]. Therefore, it is very important to make sure that the security camera can provide quality video for after-the-fact criminal investigation. However the concept of quality, as applied to surveillance applications, covers a complex range of requirements.

II.2.1 Surveillance Video Quality
The value of surveillance video as an investigative tool depends very much on the quality of the end product. Video quality however does not have the same definition for different applications.
Conventional metrics are concerned with entertainment purposes, which have significantly different requirements to surveillance. User satisfaction in surveillance video analysis mainly depends on the ability of achieving a given functionally (i.e. object recognition, event detection etc.). In contrast, entertainment-based video is regarded as high quality if the non-expert viewer has a pleasant experience watching the video [8].

Important parameters in surveillance applications include:

- Target size
- Lighting characteristics
- Impact of video compression

The importance of these factors however might be different in a range of scenarios.

The target size is the most important parameter in surveillance applications, generally referring to the percentage of the frame that the anticipated region of interest has occupied [33]. Video encoding is the second important factor by which quality of video can be degraded. Video encoding is fundamentally employed to reduce the bit rate while limiting the impact on perceived quality. Quality can be defined in terms of the extent of error introduced by the video encoder irrespective of its position in the video. However, this concept cannot fully describe the way the human visual system perceives image sequences. Perceptual quality is more dependent on the location of the error and regions containing more details will receive more attention from a viewer [34, 35].

A surveillance video with a noisy background can still be considered high quality video as long as it can provide genuine detailed information in regions of interest. In contrast, semantic content reduction in the region of interest (e.g. face, object textures etc.) not only degrades the perceived quality but also might be misinterpreted by people who are monitoring or investigating a crime scene. This is a sensitive subject as this content variation can even occur in seemingly benign video processing stages. Such errors are difficult to perceive in regular normal-speed viewing, yet still have significant impact on post-hoc human (or computer) recognition ability. In fact, the best approach to resolve this issue is arguably to design a surveillance video system in which the region of interest has been minimally impacted by different processing stages.

**II.2.2 Complex lighting conditions**

One of the main challenges in the use of surveillance systems is overcoming complex lighting conditions. Thus, a surveillance camera capable of coping with lighting changes would be very much in demand [36, 37]. While many high quality megapixel cameras are suitable for identification, they generally are not very efficient in low lighting. For instance, many installed surveillance systems in petrol stations are located outdoor and hence are likely to suffer from low light conditions at night, and saturation during the day [38]. Generally, capturing video in low light conditions places some significant constraints on video quality.
Capturing video in low light areas can cause significant distortions to important details. Furthermore, the amount of introduced noise in such circumstances is normally more than the artefacts resulting from lossy compression. The noise appearing on most digital frames is mostly due to photon noise [39]. The number of photons per second is low in such places and to have a low noise image (high signal-to-noise ratio) the exposure time needs to be increased. However, long exposure time means a lower frame rate which can lead to subject blur. Furthermore, longer shutter speed and therefore higher ISO sensitivity can also increase noise due to photodiode leakage currents [40]. Many consumer camcorders exploit different approaches (e.g. dark frame subtraction, noise removal algorithms etc.) to reduce the visual impact of noise [41]. However, such methods can degrade video quality by removing some information of potential forensic relevance. For example noise removal algorithms normally discard high frequency data in which some subject details might be included [42].

Biological vision systems on the other hand have a range of techniques for adapting to a variety of lighting conditions [9] and this feature enables animals, including humans, to see objects and their details, even in very dark, extremely bright and high contrast (combination of bright and dark) environments. Cameras are less adept, and correct exposure needs to be set for different lighting conditions. For instance, shadow (i.e. that can be seen as dramatic difference between light and dark areas) can complicate the ability of suspect recognition [43, 44]. Indoor surveillance cameras used to monitor outdoor activity also face the same challenge where faces or some details of a subject are obliterated. Although complex contrast enhancement and thresholding techniques can possibly extract some of the information from dark or shaded areas, changes in illumination can have a great impact on accuracy of information [45]. Dark regions of objects (such as eye-brows, eyes or even characters printed on a t-shirt) can lose detail.

Although correct camera settings in complex lighting conditions can noticeably improve the resulting surveillance video, some intelligent video processing is required to get the most out of it [46]. While a range of image processing algorithms have been proposed for different applications, the focus of some of the research presented here is to examine how biomimetic algorithms can be exploited to improve surveillance camera performance in variable and complex lighting conditions.

II.2.3 Surveillance Cameras Dynamic Range
Advanced surveillance systems are now capable of capturing high dynamic range (HDR) videos. Compared with traditional low dynamic range (LDR) surveillance cameras, such cameras have a dynamic range that is much closer to real world scenes. Older surveillance cameras use simple automatic exposure control systems to compress the real world high dynamic range.
Object recognition, particularly face recognition, in high dynamic range scenes has received more attention in recent years [47-51]. In order to preserve the visual brightness and object texture, it is
necessary to scale or map the captured image by which more information is obtained from the scene [52-54].

Accurate dynamic range mapping in HDR security cameras can result in appropriate brightness, sharp contrast and high visibility videos for better recognition performance.

The two most commonly used video enhancement engines in security cameras are local contrast enhancement [55, 56] and global histogram equalization [57-59]. Each of these methods has their own advantages and drawbacks and they generally meet some criteria at the expense of others.

Global histogram equalization offers low complexity and is able to retain the overall object structure, but can also remove the visibility of object textures. Contrast stretching is as limited as global histogram equalization as it is applied over the whole image. This method also can introduce some artefacts such as undesired halos in some regions. Local contrast enhancements perform better in preserving the object textures and can highlight important otherwise unobservable textures in an HDR image.

Lack of contrast (i.e. similar neighbour pixel intensities) is one of the most common shortcomings in LDR cameras. Contrast enhancement methods cannot be as effective in LDR images as many pixels have the same intensity value. Digital image experts will normally look at the image histogram to see if the histogram is shifted substantially towards absolute black or absolute white. Frames with narrow histogram width can suffer from lack of contrast in regions of interest. Histogram equalization or more advanced histogram widening techniques are the main approaches to retrieving more perceivable information from such frames [60]. However these methods are ad-hoc approaches – with different algorithms having different performance under different scenarios. It is highly desirable to have a single, efficient and low complexity algorithm to adaptively map and scale high dynamic range video which is effective over a wide range of scenarios.

The biomimetic approach is particularly useful for improving the contrast of object detail in HDR videos. The major discriminating feature of this approach to other local contrast enhancement methods is its comparatively low computational complexity. In this work, a biomimetic photoreceptor model of temporal preconditioning that is also a local contrast enhancement addresses the challenge of complex lighting conditions. This system helps to highlight important unobservable or small textures in moving objects in HDR video (e.g. small-scale text on suspect shirt in a dark environment). Such textures, which contain higher frequency details, can be alerted or removed in some video processing stages (e.g. lossy video compression). Enhancing them however can make them more resistant to the deteriorating effects caused by video processing stages.

II.2.4 Position of surveillance cameras

A camera, depending on the application, might be placed in different locations. Proximity to the illuminators can change the scenario and hence system performance.
For example, cameras that are used for material handling or quality assurance in manufacturing processes must perform well in different lighting conditions [61], otherwise incorrect decisions can lead to significant process degradation. Robust pattern recognition systems should be able to effectively compensate against severe highlight and backlight conditions.

The position of the camera is critical in most recognition systems. For example, the camera should not face the direct beams of main light sources. Although some camera advancements such as backlight compensation technology can enable viewers to see more details in complex lighting conditions [62], more commonly spatial and temporal enhancement is required to clearly identify objects by eliminating the impact of excessive light from the image or video. One example of the common problem in indoor surveillance is when the camera is pointed towards large glass windows, washing out the background in the daytime. The background details can be recovered by enhancement methods if its intensity variation (even if it is small) has been captured by the camera sensors. High dynamic range cameras for example are far better in obtaining background variation details. One solution to such issues is to make sure the majority of the lighting is coming from behind the camera. It is also better to direct lights in the same direction of the camera focus. In fact a better video and image quality will be obtained by following such precautions when using cameras for machine vision or surveillance tasks [63].

Dealing with outdoor lighting condition can be even more difficult because light source strength and its direction are always changing during the day. Similar to indoor applications, cameras should not look straight at the sun.

Proper light positioning can increase the accuracy of surveillance identification system by increasing the visual reach during periods of darkness. It also can increase illumination of an area where natural light does not reach the object of interest.

### II.3 Motion in Surveillance Video

Video camera development is tied to the way in which the human vision system interprets images and motion, including aspects such as frame rate and compression techniques. Humans have been shown to be better at identifying objects if shown in motion rather than stills. In fact 3D representation of objects can be inferred by watching a moving object, something that cannot be easily done when observing still images. Although some research has been done on this matter [64], the question of whether viewing moving objects (e.g. faces, number plates etc.) can be used to match identities has not been fully studied [65]. Pike et al. [66] concluded that there is a clear benefit to face recognition using rotating images rather than static images due to an increase in task-relevant information. Although single frame quality can be improved with conventional spatial filters (e.g. smoothing
filters) or any other de-noising techniques, some informative details can also be removed which can bring incorrect interpretation. Some information can be effectively extracted by using multiple frames instead. However, no agreement has been reached over using moving images or still images for object recognition purposes [67, 68]. One possible reason for this divergence of opinion could be the difference that lighting conditions can make to different scenarios and how 3D information of those objects is captured in such images. In other words, movement can help the observer to infer 3D structure from object motion. However, there will be issues with multiple frame demonstration and their analysis. Therefore, another solution is to make use of temporal information hidden in moving objects in one single frame. This approach is mainly inspired by the biological vision system [69]. Surveillance videos of high speed objects can bring more challenges to object recognition tasks and not all cameras can provide enough information about moving objects. One of the main challenges is obscured details of moving objects of interest. Low frame rate surveillance cameras are particularly susceptible, but higher frame rate cameras can also encounter the same issues in low light conditions (i.e. due to short exposure time). On the other hand, high resolution and high frame rate cameras require wider network bandwidth which is often impractical in surveillance applications. One approach to this problem is to design robust video compression which concentrates on moving object recognition-related details.

**II.3.1 Conventional Motion Detection**

Motion detection is an active research topic in computer vision. There is an extensive body of work in the field of detecting, recognising and tracking moving objects in surveillance video [70]. Intelligent surveillance video systems are being developed to supersede conventional passive systems [71]. The ultimate goal is to demonstrate that intelligent surveillance video can be more capable compared with human operators who are monitoring them. A very common approach in such systems is to have a motion detection stage in which regions corresponding to moving objects are segmented from the rest of an image [72]. Subsequent processing stages in intelligent surveillance systems such as object tracking and behaviour recognition are dependent on the motion detection stage [70]. Motion segmentation and object classification are sometimes included in the motion detection stage.

Surveillance cameras are generally assumed to be stationary, and in the research presented here only motion detection systems for such fixed cameras are considered. The key issue in such systems is to automatically estimate background from a dynamic sequence. However, some factors such as shadows, illumination changes or shaking branches can make such tasks difficult. Many approaches are proposed to resolve such problems. The most popular methods in the literature include adaptive Gaussian modelling, temporal median or average, and parameter estimation based on pixel processes [73-77]. More complex approaches have been proposed, such as employing a Kalman filter in Ridder et al. model or Stauffer et al. model in which a mixed Gaussian model is used for each pixel and background estimation is updated to adapt to illumination variances [77, 78].
II.3.2 Conventional Motion Segmentation

Motion segmentation is another process that can be either included in the motion detection stage or as a separate process. This stage aims at detecting regions corresponding to moving objects. Two of the most common specific moving object applications in surveillance applications are people and vehicles, and more specifically faces and licence plates. Subsequent processes such as object tracking use these moving regions in order to focus on certain areas and hence reduce computational complexity [79, 80]. Most segmentation methods use spatial or temporal information, but to date a method that can effectively use both spatial and temporal information has not been employed.

Some of the most common conventional approaches for motion segmentation are:

1. **Temporal Differencing:**

   The pixel differences between consecutive frames are calculated in temporal differencing methods. Two or three consecutive frames in an image sequence are normally used and different thresholding techniques are employed to extract moving regions. These methods are highly dynamic and generally not all relevant pixels are identified. The model proposed by Lipton et al. [81] is one such method that employs temporal differencing in real video sequences. The absolute difference between two frames (i.e. current and previous frame) is obtained and a thresholding method is used to determine changes. Connected component analysis is later used to cluster moving sections into a motion region.

2. **Background Subtraction:**

   Background subtraction is another method for motion segmentation. These methods are more suited to situations in which the background is quite static. The difference between the current frame and the reference background is firstly obtained and the moving regions are detected after using a suitable threshold function. These steps are performed on a pixel-by-pixel fashion. This method is relatively simple and easy to implement. However, it is sensitive to dynamic scene changes. The performance of such methods can be improved by a good background model [79, 82].

3. **Optical Flow:**

   In this method a vector field is used to describe the velocity of pixels in an image sequence and the characteristics of flow vectors of moving objects are used to detect moving regions. For instance, the displacement vector field is obtained in Meyer et al. model to segment the moving region [83]. The main advantage of such methods is that they can be applied even when the camera is not fully fixed. However, most optical flow methods are computationally expensive and therefore unsuitable for low cost surveillance video systems unless specialised hardware is employed. They are also susceptible to noise that makes them impracticable in low light and noisy situations. More detailed discussion of optical flow methods and their performances can be found in Barron [84].
II.3.3 Psycho-visual Motion Detection techniques
The principle objective in biological vision motion detection is to calculate self-motion (referred to as ego-motion) [85]. A biomimetic model should be reliable in a wide range of different environments. Considerable research effort has been made in last 50 years and different models have been proposed. One of the earliest motion detection models is the correlation-type motion detection system which was introduced about 50 years ago [11]. This model is known as an elementary motion detector (EMD). Non-linear correlation of the response of one detector by the delayed response of a neighbour is the main structure of this system. This approach is the dominant biomimetic model for ego motion detection not only in insects but also in all other biological vision systems including the human vision system[86]. The correlation based EMD method has shown to be very effective compared with other proposed biologically inspired motion detection models including gradient motion detector [87]. This model is described in detail in Chapter VI.

II.4 VIDEO COMPRESSION

II.4.1 Introduction
Video sequences undergo a digital representation and compression stage to meet limitations imposed by transmission bandwidth or storage capacity. It is tempting in surveillance systems to encode video using lossless representation techniques, but the very high bandwidth associated with video makes this impractical.

The compression systems used in the surveillance and security sectors are largely the common standardized models that are also used in other multimedia applications. The quality requirement of surveillance system however is somewhat different from other consumer-based applications (such as television and movie reproduction).

Video compression standards are proposed and defined by collaboration between the working group of ISO/IEC Moving Pictures Reference Group (MPEG) [88] and International Telecommunication Union (ITU) [89]. These standards are designed for a variety of applications such as broadcast quality television, video conferencing, and low data rate transmission to mobile devices. Most of these standards such as MPEG-2 [88], H.261 [90] and H.263 [91] have block-based coding as their common feature.

Object based coding is later proposed in MPEG-4. A recent H.264 standard, also called MPEG-4 AVC, offers better efficiency comparing to previous standards. Features supported by the encoder have been described in the abovementioned standards. These standards also include defining the syntax and semantics of the bit stream and give some freedom to codec designers [92]. There has been
significant development effort to adapt the existing codecs to reduce the data rate whilst retaining video quality.

As previously emphasised in II.2, video quality does not have the same meaning for different applications. Designers are more concerned about how their video compression algorithms are going to perform in entertainment applications (e.g. TV or cinema industries) than surveillance videos and analysis applications.

II.4.1.1 LOSSLESS AND LOSSY COMPRESSION

Lossless compression achieves a modest reduction in bandwidth by allocating efficient representative codes to common patterns in the data. Lossy compression, on the other hand, identifies information in the video content which does not contribute to perceptible quality and removes it from the data representation.

Lossless coding is preferred in applications in which size reduction is not the main issue and compression artefacts cannot be tolerated. Medical imaging, software, and many technical documents use this type of compression. Lossless compression can reduce data size by encoding repetitive information efficiently and since this information will be restored in the decoding process, no loss of information will occur.

Lossy compression is not reversible and therefore the lost data cannot be retrieved when it is removed. In this method, some loss of information is acceptable and removing nonessential detail from the original can lead to saving a considerable amount of data. In lossy video compression, the unnecessary information is defined by how people perceive the captured data in question. There is a trade-off between reducing size and preserving information in lossy data compression, compared with lossless compression, to offer a better compression rate. Therefore, considerable attention has been devoted to lossy compression algorithm design in the literature [93].

II.4.2 Quality of decompressed Video

Video data is discarded to satisfy network bandwidth limitations or storage capacity. If after the decompression process enough data from the original video cannot be recovered, the quality of video will decline and some observable artefacts will be introduced to the video. In other words, discarding data in lossy compression introduces unwanted visible distortion. Video compression algorithm designers are always trying to minimize the perceptive artefacts introduced by encoding systems. Computational complexity must also be considered to reduce the costs of specialised hardware, to meet real-time constraints, or to reduce power requirements. Such constraints mean that optimisation is focussed on the intended market for the camera, and there is little opportunity for an algorithm to be intelligent enough to adapt to different applications and environments [2].
II.4.2.1 SPATIAL AND TEMPORAL VIDEO COMPRESSION

A video encoder is a virtual engine in which compression or decompression of digital video is performed. Digital video consists of sequences of images (also known as “frames”) and these images are presented at small time intervals to preserve continuity. The human eye perceives the frames as motion instead of standalone still images. Video compression algorithms try to minimize the amount of information needed to code each still image [2]. The first stage is to compress an individual image and use information contained within that current frame using spatial information. This stage is known as Intra-frame coding and such encoded frames are called I-frames. The most widely used Intra-frame coding is based on the JPEG method in which the Discrete Cosine Transform (DCT) is used to reduce the accuracy of high-frequency spatial components or remove them altogether, which are generally not perceptible by the human psycho-visual system. Temporal redundancy uses a technique known as block-based motion compensated prediction, which exploits the fact that frames do not vary very much from one to the next. A block of pixels is matched to a similar pattern in a reference frame, and represented by a vector offset and, if necessary, by a block difference correction. Temporal compression or inter-frame coding was first introduced in MPEG-1 standard and later refined in MPEG-2 and H.264. Predictive coded frames (P-frames) and bidirectional predictive coded (B-frames) are the main types of inter-frames. P-frames only rely on previously transmitted frames and use their data with minor changes. B-frames however can also interpolate between past and future reference frames. Motion vectors indicate the number of pixels by which the macro-block has been shifted from previous (or succeeding) frames. Using motion vectors and finding the best match macro-block will still result in some prediction error. The prediction error is also DCT transformed, quantized and added to the encoded bit stream (using variable length coding). The video is reconstructed by reverse operations.

Intra- and Inter-coded frames have their own advantages and disadvantages. Inter-coded frames, compared to intra-coded frames, have fewer bits per frame and using them in encoding video will result in lower bit rate video. However, Intra- coded frames will let us to have more access to the sequence at all times and that accessibility is arguably crucial in surveillance applications. Intra-coded frames also prevent subsequent quality degradation. Such quality degradation occurs in Inter-coded frames due to errors propagating from one Intra- coded frame to another [19, 94].

II.4.2.2 ARTEFACTS IN COMPRESSED VIDEO

Block based coding is the one of the most popular approach for video coding. Most of the current video coding standards (e.g. MPEG-1/2/4 and H.26x) use the Discrete Cosine Transform (DCT). DCT coefficients of non-overlapping blocks are calculated and then quantized in the encoder. In the decoder, de-quantization and inverse transform is performed to recover the original data. Heavy truncation of high frequency DCT coefficients leads to a constant value for many elements of
CHAPTER II CONVENTIONAL SURVEILLANCE VIDEO TECHNOLOGY

compressed macro-blocks, visible as blockiness. Discontinuity between adjacent blocks will result in such artefacts and is a main consequence of independent quantization of DCT transformed coefficients [95-97]. The larger the quantization values the more data that will be discarded and hence the lower the data rate video will be generated at the encoder. Therefore, more noticeable artefacts will be introduced in the decoded video.

In video sequences, block edges become visible in smooth regions of the image, while areas of high spatial frequency detail effectively hide the inaccuracies introduced by such compression. However such errors can introduce misleading details when captured surveillance videos are used for further investigations.

For example, compression can have a destructive impact on the recognition of small objects. The licence plate shown in Figure II-2 is hardly recognizable due to artefacts, including blockiness, introduced by the encoding process.

To avoid such cases, high spatial frequency details in the region of interest should not be coarsely quantized.

Different quality metrics for blocking artefacts have been proposed in the literature [96, 98-100]. The proposed artefact measures are either used as a standalone quality metric or a factor in quality evaluation. Referenced and unreferenced objective video quality metrics have been developed.

The referenced approach requires access to the original video sequence to generate a metric based on differences between the original and compressed video frames [101]. Unreferenced approaches are the

Figure II-2 this picture demonstrates how video encoder can impact on license plate recognition which has been problematic in many criminal investigations (In this example video is encoded by Motion JPEG compressor with quantization scale of 17)
only option when a reference is not available, and specifically for assessing the quality of a stand-alone compressed video sequence [95, 96].

Since our understanding of the biological eye is somewhat limited, designing such quality metrics is a very difficult task. Effective unreferenced objective quality metrics can only be designed if at least some prior knowledge about the distortion type is available to the designer [97].

II.5 GENERAL CHALLENGES IN FORENSIC VIDEO ANALYSIS

It is common for ad-hoc approaches to be applied to image and video enhancement for forensic analysis. The difficulty is that standard procedures are poorly defined and to the extent that such protocols exist, they are quickly outdated by new technology developments. In fact, forensic analysis professionals are often not involved in the development of such protocols [102].

The main issue is that there is usually a reasonable doubt in identifying the suspect or visual event especially if the equipment is of poor quality and the personal judgement of the jury determines the acceptance of any video material presented as evidence in the courts [103]. There is an argument that the acceptance or rejection of video evidence should be based on a forensic expert’s scientific analysis [1]. Well studied scientific approaches need to be set and applied to the enhancement of video captured by surveillance cameras. This ensures that the optimal techniques are implemented to minimise the deteriorating effects of capturing devices without having a significant impact on useful information. This clearly involves knowledgeable qualified experts to use accurately calibrated procedures to process the video and verify that the process has been applied as appropriate. Due to the expertise required to carry out the video enhancement process, in-house technicians might not be the best people to perform these complex technical tasks. Law enforcement authorities should be aware of the technical structure and the complexity of the systems used to record, process and enhance the video [104].

II.5.1 Surveillance Video quality in forensic analysis

Forensic video analysis is the relationship between video and law enforcement. It is a relatively new scientific tool for enhancement and identification of objects and individuals in recorded video crime scenes.

A short surveillance video of a person of interest may be released to the media in order to obtain more information from the general public. However, often there is a lack of details in regions of interest [1]. Releasing such poor quality videos will obviously not describe any distinctive features of the suspect to people who are attempting to recognize that person. The face and many of the details of clothing of the suspect are a blur.
CHAPTER II CONVENTIONAL SURVEILLANCE VIDEO TECHNOLOGY

Digital evidence can be excluded in courts if its weakness is successfully challenged by defence counsel. The current judicial trend in admission of CCTV video footage shows that poor quality videos have a strong chance of being excluded and the case is likely to be lost if this is the only key evidence [1].

Surveillance footage might be used for different purposes and different observed activities, which can be subject to scrutiny after the fact. If the surveillance system is implemented to monitor and control an area, the user is only able to monitor the speed and direction of an object (if the presence of such an object is known). However still images obtained from surveillance video have long been used as evidence of actions in a range of applications (e.g. criminal activities, accident investigations etc.) [105].

The recognition of distinguishing features of suspicious objects has been challenging for investigators in these circumstances. Such identification tasks cannot be performed unless cameras capture an adequate amount of information. At times, the user only needs to detect the suspicious object and ascertain with a high degree of certainty whether the object is present. In such cases, the object mostly occupies a considerable portion of the screen area. By increasing the percentage of the area occupied by the object, it is possible to recognize and identify the object (for example a crime suspect). It is not an aim of this work to suggest any appropriate image size to meet any particular surveillance systems requirement and there is no guarantee that objects will be identifiable because their size is large in the screen. In fact, other factors including lighting, angle of view, camera quality factors etc. can have a significant influence on recognition accuracy.

CCTV systems commonly rely on recording technology and video footage is saved onto digital storage media. Because such systems have a finite storage capacity, they can only retain videos for a set period. Such data will be overwritten after couple of weeks. A retention time of roughly 31 days has been recommended for CCTV applications [106]. This period is supposed to provide sufficient time for authorities to retrieve the video in the event of serious incident. However there is no guarantee that this data is only needed for a short period of time. On the other hand, adding additional facilities to store more video data can result in substantial cost for the owner. Compressing digital images is the leading solution to save more data on the hard drives. This compression will invariably reduce the quality of video. It is therefore vital to inspect the quality of compressed recorded video to make sure that no substantial difference has been made to the original video. The artefacts introduced by sophisticated processing, such as compression, can complicate the scenario by discarding required information [2]. In fact, such unwanted artefacts can threaten the usefulness and fidelity of such evidence [3]. In other words, identification experts may be misinformed due to the nature of the representation and hence subjective opinions of the evidence usefulness may not be adequate for legal application. Forensic investigators who confront this issue every day assert that having less artefact in
processed video can help them to validate obtained evidence in terms of reliability and admissibility. This type of evidence is frequently challenged in courts [107].

II.5.2 Video compression and forensic analysis

In lossy compression much better reduction in file size is achieved. However, hardly noticeable information (that is not always “irrelevant”) is discarded.

It is common for digital video to contain visible artefacts due to lossy compression, such as blurriness, blockiness, smearing and repetition of objects. Whether or not there is subsequent editing and recompression of the video, the reliability of any digital video presented in court has been questioned due to the potential for content alteration during compression stages [108]. Video might be modified intentionally, innocently, or by the inadvertent introduction of compression artefacts. It has been demonstrated that lossy compression alone can be the main source of information alteration [4]. As previously mentioned compression is not reversible and cannot be retrieved in the decoding process. Visible artefacts caused by lossy compression can raise questions of the validity and therefore admissibility of surveillance video as it is an increasingly common form of legal evidence. The main issue here is that the decision to discard data is only determined by the compression algorithm and it does not make any assumption about usefulness of such information in forensic studies. The higher compression ratio used, the more information will be lost in resulting images and, depending on the application, the image will be less useful for further investigations. The compression parameters are mostly set by camera firmware and processing software. Capturing and encoding systems are like black boxes to end users who normally have little access to compression parameters. Many of the artefacts generated by compression algorithms are unavoidable. Therefore, it is highly desirable to devise new surveillance systems to make sure that important image content is not lost or obscured.

II.5.3 Conventional techniques in forensic analysis

In forensic cases involving pattern recognition tasks (e.g. face or license plate recognition), there is normally no control over the quality of the video captured by a CCTV camera at a crime scene and it is often of very low quality [109]. Many forensic experts will make use of image processing techniques to extract more information from obtained surveillance footages [110]. One common approach is to use different spatial and temporal filtering, of which linear filtering has been widely used in many investigative scenarios. Edge enhancement, deblurring, and sharpening are mainly included in linear filtering [111]. They are mostly aimed at contrast enhancement of small detail in an image. One issue with this type of enhancement is that the image will no longer remain an accurate representation of the scene if a high degree of such enhancement is used, although they still might be useful as an adjunct for detail interpretation in court. Another downside of using a high degree of such linear enhancements is noise amplification that also can distort some important information. Non-linear contrast enhancements including gamma correction, grayscale transformations and look-up
tables are also implemented in many forensic studies [112]. Brightness ranges within frames is normally selected and contrast is adjusted. Severe contrast adjustment can cause loss of detail and artefacts. Noise reduction techniques are also applied to many videos and images to reduce their deteriorating effects on semantic information [113]. Similarly, overuse of this technique can result in removal of relevant detail. Different types of filters (e.g. blurring, median and despeckling etc.) can be used in order to suppress noises. One important caution that needs to be taken is to document all of the applied enhancement steps. This is to enable other individuals to follows these steps and produce the same result when the image is subjected to forensic analysis.

One standard routine in forensic image analysis is image restoration in which an image degraded by a known cause (e.g. motion blur, optic blur etc.) will be processed through an inverse filter in order to retrieve more information. Quite obviously there are limitations imposed by introduced noise and the degree of the information that is lost. Information that has been totally lost cannot be replaced. Blur removal is one such technique. Such blurring phenomena can be described mathematically and the inverse filter might be able to compensate for its effects. In Figure II-3 we see an example in which some de-blurring techniques were used to restore some lost information [114].

![Figure II-3](image)

*Figure II-3* frame 1: a motion-blurred video- frame 2: magnified view- frame 3: magnified view using a de-blurring technique (this picture is from Cho, S. et al [114])

### II.6 Summary

A video surveillance camera is a device used to capture information. However, we need to define what it is that we want to capture and what level of detail is required for later investigation. One possible suggestion is that collecting more pixels will result in better video quality and hence more details. However, capturing more pixels can lead to wasteful systems design of storage devices, network bandwidth and all related infrastructure costs. If our goal is to improve the video quality captured, we need to consider all different scenarios in which the video quality can be impacted.
including the lighting conditions, video processing stages, camera placement and all other relevant factors.
In this research we investigate the root causes of common issues and modify the surveillance video processing by which we can improve the collection of useful video content.
This chapter is focused on biological vision systems with a specific interest in insect vision systems. The similarities and differences with the human vision system will also be discussed. Human eyes have some similarities to a camera (a highly simplified diagram of eyes [115] is shown in Figure III-1). The complexities of the human eye however far surpass the capabilities of a digital camera system. The human eye is able to look at a scene and dynamically adjust its system parameters [116], enabling the perception of objects from different distances and in a variety of lighting conditions [117]. In other words, what is seen is not the actual light received by the eyes, but a reconstruction of objects based on this provided input. The opening and closing function of the pupil is set by different perceived region brightness. Although we may not be able to fairly compare the camera with the human eye, roughly speaking, human eyes have a range exceeding 30 f-stops [118]. It should be noted that the f-stop range only depends on the lighting conditions. However, the human eye can also adjust itself to different subject contrast, which is not the case in cameras.
Overall, the human brain is able to intelligently interpret information from the eyes [119], whereas with the camera, the raw data captured by the camera sensor is processed using relatively simple biomimetic models to produce an image acceptable for interpretation by the human visual system.

### III.1 Human Vision

In order to understand how the human vision system works we need to know more about the some of its physical attributes. The interior chamber in human eye (i.e. located between the lens and cornea) is filled with the aqueous humor that is a transparent fluid similar to plasma and the posterior chamber also (i.e. located behind the lens) is filled with another type of liquid (i.e. vitreous humor) which is more gelatinous. The primary refractive power is sourced from the cornea optical interface with help of an internal lens by which human eyes are able to focus on close objects. A relatively large aperture in the camera-like human eye resulting from using such refractive power allows good light gathering and will provide good static acuity specifically when the object of interest is located at a short focal distance. Camera sensors typically use a single large aperture to mimic the way that the cornea and lens perform in the human eye.

The retina is composed of many layers of neural tissue. This tissue contains the photoreceptors which include rods, cones and non-imaging photosensitive ganglion cells. The main photoreceptor cells are rods and cones, distributed differently across the retina. The rods are normally smaller and also more sensitive to smaller amounts of light than cones [120]. The ganglion cells do not contribute directly to sight, but are believed to support pupillary reflex [121].

Cones need significant light to produce a signal. There are three different types of cone cells that are categorised by their response to different wavelengths of light. Colour vision is the result of these three types of cone cells operations. At very low light levels however the visual experience is based solely on the rod signal. This explains why colours cannot be seen at low light levels as only one type of photoreceptor cell is enabled. Table III-1 summarises the properties of rod and cone sensors [120].
### Properties of Rod and Cone Systems

<table>
<thead>
<tr>
<th>Rods</th>
<th>Cones</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slow response: long integration time</td>
<td>Fast response: short integration time</td>
</tr>
<tr>
<td>High amplification</td>
<td>Less amplification</td>
</tr>
<tr>
<td>High sensitivity</td>
<td>Lower absolute sensitivity</td>
</tr>
<tr>
<td>Low acuity</td>
<td>High Acuity</td>
</tr>
<tr>
<td>Saturating Response</td>
<td>Non-saturating response</td>
</tr>
</tbody>
</table>

The physiology of human colour perception has been intelligently employed in video technology. For example, the human vision system is more sensitive to variation in brightness than colour, which is why a video system can devote more bandwidth to the luminance component (Y) and the rest of the bandwidth will be allocated to the colour differences components (i.e. chrominance blue (Cb) and Chrominance red (Cr)). This video colour optimization method does not have a significant impact on the human visual experience. The other feature of human vision system, employed in video compression, is its ability to see brightness variation within different spatial frequency ranges. Humans are fairly good at observing small variation over a large area, but are not very competent at perceiving the detail of rapidly varying brightness variation over a small area (i.e. high spatial frequency components). Data can be compressed by quantising high spatial frequency components to low accuracy [2].

These techniques suggest that many of ideas employed in current video processing have been inspired by human perception physiology (see Figure III-2 and Figure III-3). However, there is still a long way to go to deeply understand how the human vision system functions in different conditions and how we can utilize its features in real applications. There are other biological systems that offer less complexity and therefore are much easier to understand. Many of these systems can still be very beneficial to real applications such as video enhancement and robotics. The insect vision system, for example, has an outstanding ability to adapt to different lighting condition in order to enhance and encode the information captured by the eye’s lens [122-125]. Insects have a great ability to detect and follow moving objects in complex situations with much lower computational power in comparison...
CHAPTER III – BIOLOGICAL VISION SYSTEM

with the human vision system [126]. This feature makes this system an attractive research area and contributes to understanding the human vision system which has many features in common [127].

Figure III-2 this block diagram demonstrates how conventional video compression algorithms can compress video by discarding the data that is hardly visible to human vision system In other words, human vision features are exploited to save some space (or bandwidth). Discarded data however can contain some vital information that can be useful in surveillance applications. The focus of this research is mainly to employ Insect vision system model in spatial and temporal compression component of video compression in order to retain abovementioned vital information.

Note: Color compression will not be discussed in this research work.
In fact, many of techniques used in current video technology are inspired by human vision system. The lower table suggests how we can similarly pursue advancing surveillance video technology by employing different components of insect vision model based on their features.

**Note:** *HVS:* Human Vision System  
*IVM:* Insect Vision Model

---

### Figure III-3 Video technology development

The cinematography technology development history leading to current digital video technology.
III.2 INSECT VISION SYSTEMS

There has been a growing interest in understanding the principles of the biological vision system and applying these to the design of new image capturing devices for a variety of applications in the last two decades. This section is aimed at clarifying some concepts in the field of insect vision and flight guidance that will be used later in our experiments. Insects use vision to stabilize flight and avoid collisions with surrounding objects. They are also able to navigate to a far destination to locate food sources. Their ability to achieve a smooth landing has also been a compelling topic of interest for scientists and many insights from insect vision systems are now being used to develop new guidance system in terrestrial vehicles. This section describes some of the outstanding features of the insect vision system, particularly in the area of motion sensing.

The arthropod (e.g., insects, crustaceans) eye structure is somewhat different to the human eye (i.e. vertebrate eye). Their components are arranged differently. In an insect eye, a single ‘facet’ and a single lens covers a retina of many sensory cells – an ommatidium in which 7 to 11 sensory cells are included. In other words, each sensory cell does not contribute one ‘pixel’ to the final image like the human eye (see Figure III-4 – this figure is obtained from [128]).

The number of sensory cells per square millimetre that are packed in a hexagonal array in human eye is very different to the number of sensory cells in compound eyes (i.e. 175000 in a human eye vs. 30000 in a compound eye). The most sensitive region in the human retina (i.e. fovea) is where these cells are concentrated. These numbers vary significantly between different arthropods. For example, the wingless silverfish have only few ommatidia, while the dragon fly has 30000 ommatidia in each eye. It can be concluded that human eyes are able to detect greater spatial resolution due to the large difference between its sensory cells versus insect eye.

The human eye works better in illuminated areas and although it has a certain degree of night vision, it is not very efficient in detecting details in darker areas. The human optical system requires more light to be able to see objects clearly. Many insects struggle to see in low light conditions. They are limited by the small apertures of each ommatidium in the compound eye. Therefore they are not very capable of detecting contrast in dim lighting. However some insects such as dragonflies will have better performance with the help of their apposition eyes with wider facets. They can collect more light over a longer period before integrating the signal to produce a final result [129].

The insect vision system responds far better to moving targets than stationary ones - it is excellent at detecting motion. This explains why honeybees, for example, are able to see wind-blown flowers more readily than still ones. The “flicker” effect in insect eyes occurs when an object moves across the visual field and ommatidia will be turned on and off progressively.
Despite having low resolution eyes, insects in particular are very good at perceiving motion in a moving cluttered background. In contrast to mammals, the insect visual system has no optical zoom and hence cannot change focus. This feature of insect vision can be similar to the camera in which its focus cannot be decided efficiently unless it is manually controlled.

Such similarities inspire the innovative ideas of applying insect vision system in surveillance cameras to efficiently encode data and detect motions in complicated lighting conditions.

**Figure III-4 Structure of Building Block of the fly’s retina**

Insect eyes are composed of units called “ommatidium” which contains a cluster of photoreceptor cells. Each ommatidium appears on the surface as a single polygon, called a facet (this figure is obtained from [128]).

When light enters this ommatidium from above, through the facet, the cornea and crystalline cone together focus the light onto the visual cells. This system forms the light-focusing of the ommatidium.
The insect vision system consists of different stages [10] of which three main stages are studied and implemented in this work (see Figure III-5):

- **Photoreceptor cells**
- **Laminar Monopolar cells (LMCs)**
- **Elementary Motion Detectors (EMDs)**

Comparing this system with the human vision system, we can see that the photoreceptor stage is fairly similar to human photoreceptor stage. However, LMCs are only identified in insects and EMDs are perceived as a processing stage rather than biological cells [10].

---

*Figure III-5 – Overview of insect vision stages studied in this research.*

*Each of these block diagrams are described later in Chapter IV (Photoreceptor), Chapter V (LMCs), Chapter VI (EMDs) These stages are the system models that are used in our study, to mimic the real life insect vision system.*
III.2.1 Photoreceptor (phototransduction) Features

Photoreceptors are the cells in the retina that respond to light. Their distinguishing feature is their ability to efficiently encode high dynamic range luminance intensities so that more object features can be retained in both light and dark areas. The encoded information in this stage will be sent to the brain via a bandwidth-limited neural network. The photoreceptor uses temporal information, and different non-linear temporal filters are employed in this biological vision system component.

The biological vision system’s extremely versatile light detectors often outperform artificial devices. For example, a photoreceptor cell in human eye can respond to just one photon [130]. In fact, only the most sensitive advanced detectors can compete with biological vision systems. Studying these photoreceptors could, it is argued, lead to the development of “bio-quantum” devices in which biological components and man-made technology are combined in order to improve the capability of existing systems [131].

In this work, we study and implement the nonlinear temporal processing performed in photoreceptor cells which are inspired by blowfly phototransduction (see IV.2.2). This is a modified version of the photoreceptor model, which was initially proposed by Van Hateren and Snippe [132]. This model responds to change of intensity rather than intensity itself, reducing the impact of background intensity in later stages.

The main task of this temporal processing component is the dynamic correction of pixel intensities based on lighting conditions. Pixel values are dynamically controlled by several parameters including the corner frequency of a low pass temporal filter and non-linear saturation transform (i.e. using Naka-Rushton (see IV.5.3 for more details). In other words, pixels range is efficiently compressed by amplifying gain in dark areas, while simultaneously reducing this gain in high luminance areas.

The temporal filtering in the photoreceptor enhances spatial details in processed video. In simple terms, video can be spatially enhanced by using temporal information [21]. The major drawback of the photoreceptor model is that it does not enhance stationary objects and it may have a brightness reversal problem. For instance, some of the transformed intensities might fall outside the normal range pixel range (e.g. 0 to 255 for 8 bit dynamic range) due to non-linearity and need to be manipulated in order to be compatible with current display technologies. In comparison with typical contrast enhancement approaches (i.e. gamma correction, histogram equalisation etc.), this system considers the local property of a pixel intensity and that is why object textures in poor illumination can be recovered very well in a biological local contrast enhancement system. It also needs previous frame pixel intensity to enhance moving object features. This method also requires considerable calibration in order to achieve significant results. Since this method has been newly implemented, it has to be tested in different scenarios with different lighting condition to find the most optimised system.
III.2.2 Laminar Monopolar Cells (LMCs) Features

This stage mimics the spatial-temporal processing performed in flies’ second-order neurons. This stage is very similar to bipolar cells in mammalian eyes [125]. Space and time redundancy is removed by which motion detection tasks in the next stage (Motion Detection stage) can be performed in an information theoretic optimal way [133]. Most of the DC component of the input signal is blocked in this stage which helps to discard the redundant details in both the space and time domain in an efficient way.

Parameters in LMC filters are varied in both space and time depending on light levels. The redundant part of the signal needs to be neglected; otherwise, it can have detrimental effects on system efficiency. For example, it can introduce delay and errors to the designed system. Therefore this stage introduces additional neural processing, inspired by lamina monopolar cells (LMCs) in flies, that is employed to remove spatial and temporal redundancy in the most optimal way based on local light levels [133].

Variable temporal filtering and relaxed spatial high pass (first-order) filtering and a non-linearity saturation transform (i.e. using tanh) are included in this stage (see Figure III-6 - taken from from Juusola, M et al. [134]). An object’s structural details are propagated to the next stage only if there is motion in that area, otherwise it will fade after a few seconds (depending on temporal filtering parameters). In other words, the large lamina cells respond to the temporal derivative of the intensity with maximum slope of the response curve at the most common level of contrast. This spatiotemporal processing removes the temporal correlation introduced by the duration and shape of the impulse response of the photoreceptor [135].

The model used in this research work has some differences to the model proposed in the literature. The inversion of sign observed in neurophysiological recordings from these cells is not included as it does not have any influence on the performance of the system [136]. Similarly neural superposition is not included. Neural superposition involves the combination of a number of independent samples of the same point in space to reduce noise [136].

The inclusion of LMC processing to the third element (motion estimation stage) provides significant benefits for motion velocity estimation [10]. The LMC model with some modification has been used as a biologically inspired contrast enhancement in Chapter V. The enhanced image resulting from this model is significantly different from other conventional contrast enhancement methods. In fact, we need a contrast enhancement method that can efficiently amplify moving object structural details and can also resist the noise and artefacts introduced by different video processing stages including video compression. The LMC contrast enhancement has the ability to not only enhance moving structural detail but also create a distinctive border around these details that makes them less susceptible to video compression degradation. In simple terms, if we assume that conventional contrast enhancement increases the distance between all high and low luminance pixel values, the LMC model
only increases the distance between pixels that are positioned around the border of moving object structures.

### III.2.3 Elementary Motion Detectors (EMDs) features

Motion detection is a critical aspect of an insect vision system. The correlation detector model has been widely used to explain the direction-selective motion detection system of insects. The main feature of insect vision is the ability to operate effectively in complex lighting conditions, particularly under low signal-to-noise ratio [137]. In this mode, each subunit processes motion by multiplying the output signals of two neighbouring pixels after one of them has been low-pass filtered. The output signals of both subunits are finally subtracted to enhance direction-selectivity. Two separate inputs are required to infer the direction of moving object. Correlation methods are normally hypothesised as a model that is relatively insensitive to noise [138-140]. The EMD-based models however have to be fully investigated to see if they can perform robustly across different lighting conditions.

This model consists of two symmetrical subunits (i.e. so-called half-detectors).

Balanced correlation models can perform better in comparison with imbalanced models and that is why half-detectors are not as popular in animal vision models. The balanced correlation model’s output is generated by subtracting the response of one half-detector from the response of the other symmetrical half detector (see VI.2 for more details).

As its name suggests, this element (local motion estimation) is the core of our system by which motion is estimated in surveillance videos. The other two preceding elements are pre-processing spatial-temporal stages that improve the motion estimation performance. This stage is an elaborated version of the Elementary Motion Detector (EMD) proposed and developed by Hassenstein and Reichardt [141].

The elaborations to the main EMD are spatial high-pass filtering (employed before EMD) and applied non-linearity saturation transform (employed after EMD). Saturation non-linearity transform mimics the limited bandwidth of neurons in which some of the transcoded values need to be bounded in order
to be propagated to other stages. The gain parameters are normally estimated to make better use of available bandwidth [10].

Applications of the photoreceptor, LMC and EMD are demonstrated in next three chapters. These experiments are specifically tailored to show how each biological vision system component can help to improve surveillance video quality. For the purposes of clarification however, we briefly look at the research flow to produce seemingly smooth transition from discussion and literature review chapters to experimental chapters.

### III.3 Research Workflow:

Each of the first two elements (i.e. phototransduction and spatial-temporal redundancy reduction) has its own features and hence is individually beneficial to the motion estimation element (i.e. EMD). The first element of this model is purely temporal and the second element (Spatial-Temporal redundancy reduction stage) is a combination of spatial and temporal processing. In order to study their important features, they are individually studied and have been tested in different scenarios based on their features. Although these are pre-processing spatial-temporal stages to improve the motion estimation performance, they can individually be employed in a range of applications.

In the first experiment, we made use of the photoreceptor model to determine if there is any advantage of using this model as a pre-processing stage by which pattern recognition system performance can be improved. A simple yet robust card suit recognition algorithm was designed and implemented to study the impacts of lighting conditions on pattern recognition algorithms. The efficacy of the designed pattern recognition algorithm was firstly determined in an ideal noise free synthetic video. This experiment demonstrated that the recognition system can recognise card suits perfectly under ideal conditions. We then examined the system in unrestricted conditions in which the environment, lights, and video quality are not optimal. The performance of the pattern recognition system degraded significantly in such non-ideal conditions. Using the photoreceptor model as a pre-processing stage considerably alleviated the deteriorating effects of lighting.

Our next experiment was quite different to the first stage. The main purpose of the LMC is to remove spatiotemporal redundancy. A second experiment was conducted to examine lossy video compression impacts on alpha-numeral character recognition systems. The LMC model was employed to enhance moving structural details and mitigate possible negative impacts of compression on subjective and objective character recognition. This model was initially tried after the video encoder as a post-processing stage. Preliminary results however were not promising. It appears that distortions introduced by the video encoder were amplified in this (post-processing) stage, cancelling out benefits obtained from enhanced structural details. In contrast, results obtained from using LMC as a pre-processing stage were favourable and significant improvements were found. Various lighting
conditions were examined by using a synthetic video under which a range of scenarios was simulated. In fact, we were not able to separately study lighting conditions by using real videos. This is not only because testing different lighting condition would be costly and time consuming, but also capturing a noise free video in low light condition would be almost impossible with current capturing devices.

In the third experiment, the second and third elements of the insect vision model were used in surveillance video coding as a pre-processing stage for elementary motion detection. The first insect vision system element (i.e. photoreceptor model) was initially included in our proposed pre-processing stage but since a low dynamic-range video was used in this experiment, this element was removed to reduce system complexity. The photoreceptor model can perform considerably better when high dynamic range video is used. Therefore, the collaboration of LMC and EMD separates the moving objects from background in order to facilitate different strategies for encoding each classified areas. Region of interest (ROI) coding employs an encoding algorithm by which the data can be compressed with less effect on the viewer quality experience. For example, higher quality within the ROI normally gives the viewer an impression of high quality video without changing the total amount of data. We therefore considered how elementary motion detection can inform the video coder to reduce data rates while retaining structural details in moving objects. This study can help surveillance video compression designers to improve their system by choosing more effective video encoding parameters.

One of the most challenging problems in surveillance videos is detecting and recognising individual faces in cluttered scenes with complicated lighting conditions [142, 143]. We considered several issues that can potentially have significant impacts in recognising a person in a surveillance video footage. We also conducted a preliminary study on how our proposed biologically inspired video encoder can improve automatic face recognition systems (see Chapter VII). This chapter demonstrates that video processing techniques can be improved by employing biological algorithms, justifying further research bridging the gap between biological vision studies and surveillance system.
Chapter IV  PHOTORECEPTOR AND COMPLEX LIGHTING

IV.1 PROBLEM STATEMENT

Lighting is an essential factor in any surveillance system. Object recognition, whether it is performed by machine or human, needs to have adequate lighting to make the objects of interest visible. Many factors including the types and positioning of light sources can play an important role in lighting conditions. Each surveillance system faces its own particular challenges based on physical layout, camera features and atmospheric conditions.

Compared with simple biological eyes, human eyes have a better sense of differences in brightness between one object and another, referred as “contrast perception”. Our eyes’ contrast sensitivity enables us to see in detail what is around us. Contrast is usually defined in terms of the ratio of light intensity in one area compared to the intensity in another.

In this chapter we consider the impact of lighting conditions on object perception. A biologically inspired video enhancement is described and implemented in an elementary application-tailored object recognition system. It is demonstrated that insect vision system temporal pre-conditioning (i.e. using a photoreceptor model) can be employed to improve the recognition accuracy when lighting conditions are not ideal.

Pattern recognition algorithms need to operate under a range of different lighting conditions. There are several methods by which this can be achieved to alleviate deteriorating system performance.

IV.2 LIGHTING CONDITIONS AND CONVENTIONAL VIDEO ENHANCEMENT

Video enhancement can be one of the most important components of security video technology, especially noting security applications at night time. However, it is common for the scene context to be either hidden or contaminated with noise due to poor illumination. A large number of methods have been proposed to address this problem.

There has been a lot of research in the area of video and image enhancement over the last 40 years [144] and there is still significant room for improving the quality of the enhanced video subject to computational complexity. There are slow methods (i.e. methods using frequency domain such as Fourier transform methods etc.) that produce outstanding enhancements but which are too complex to be implemented in security cameras. Therefore, faster methods generating videos with medium enhancement is of more value in real time applications. With this in mind, we consider for comparison only conventional spatial domain methods that are used in current security cameras. Conventional video enhancement techniques are not able to extract any high quality background
information. Low quality video under low light are sometimes so dark that all the information is already lost in those regions and no matter how much contrast enhancement you apply, lost information cannot be restored. There are also video enhancements methods which fuse information from other frames [145]. This method requires large memory to store several frames and is also computationally expensive. The other downside is the ambiguity in deciding how to pick the high quality features while keeping all the low quality important information.

**IV.2.1 Histogram equalization**

Histogram equalization (HE) is one of the most commonly used methods for contrast enhancement. In this method, the pixel intensity histogram of image is altered to become closely matched with a uniform distribution.

The HE algorithm is described as follows:

Having the input frame $Y(i,j)$ that has total number of pixels $M$ and $K$ luminance levels, we firstly find the probability density function ($PDF$) of the input image:

$$p(k) = \frac{m_k}{M}, \quad \text{for} \quad k = 0,1,\ldots,k-1 \quad (IV-1)$$

We then calculate the cumulative distribution function ($CDF$) of the input image,

$$c(k) = \sum_{n=0}^{k} p(n), \quad \text{for} \quad k = 0,1,\ldots,k-1 \quad (IV-2)$$

And the CDF values are used to remap the luminance levels of the input image,

$$y(i,j) = c(Y(i,j))$$

$$\hat{Y}(i,j) = y(i,j) \cdot \{\max(Y(i,j)) - \min(Y(i,j))\} + \min(Y(i,j)) \quad (IV-3)$$

There are two main issues in the application of HE method in videos with complex lighting condition. The first disadvantage of HE is its poor performance in retaining local detail due to its global frame treatment. In fact, small scale details, that are associated with small bins of histogram, are often lost. In practice, we see some parts of the resultant frames are washed out as will be shown in the simulation section.

The second disadvantage is that the whole frame needs to be stored in order to generate the probability density function even before starting the equalization limiting real time implementation especially if high resolution frames are used.

**IV.2.2 Tone mapping**

Most LCD and related displays only support a low dynamic range, hence there is a need to compress the wide dynamic range for display [146]. Tone mapping is a common technique used in image processing to map one set of intensities to another and can represent high dynamic range images in
media with a more limited dynamic range. This technique is performed in the luminance channel only, on a logarithmic scale.

There are two types of tone mapping techniques proposed in the literature: global tone mapping and local tone mapping. In simple terms, global tone mapping is independent of local spatial context and the same operation is performed on each pixel. Local tone mapping adjusts each pixel in the context of its surroundings. Although global tone mapping can lead to unsatisfactory results (specifically when there is a large local illumination variation), it is more common in current media technology due its lower computational complexity. One such tone mapping, proposed for use in night vision, is based on logarithmic µ-law voice telephone compression [41]. The mapping function is:

\[
y' = 255 * \frac{\log(\frac{y}{255})(\Phi - 1) + 1}{\log(\Phi)}
\]  

(IV-4)

In this equation \( y \) is the pixel intensity of the original frame and \( y' \) is the pixel value of the enhanced frame. The shape of the correction curve can be controlled by \( \Phi \). This equation is similar to the traditional gamma correction function except that author claims that it works better in complex lighting conditions. Gamma correction can be simply defined by the following expression:

\[
y' = A \cdot y^\gamma
\]  

(IV-5)

Where \( A \) is the constant (in the common case \( A = 1 \)) and \( y \) and \( y' \) are input and output pixel values

**IV.3 BIOLISTICALLY INSPIRED CONTRAST ENHANCEMENT**

We propose a computational model of photoreceptor behaviour to address the contrast reproduction problem and also used as an aid for an automatic object recognition system. The aim of this work is to provide a novel temporal video enhancement method that produces contrast-rich results that are useful in practical settings specifically in surveillance videos.

In addition, the biomimetic model operates independently on each luminance pixel. Contrary to conventional contrast enhancement methods, our proposed method only responds to change of intensity (not the intensity itself). Pixel intensities are efficiently enhanced by amplifying gain in dark areas and reducing the gain in high luminance areas.
**Chapter IV – Photoreceptor and Complex Lighting**

**Figure IV-1 Conventional Methods vs photoreceptor**

- First column (from left) is the original and enhanced frames.
- Second column shows the magnified view of character label.
- Third column is the histogram of each frame.

Most of the characters have become obscured after conventional contrast enhancement methods (i.e. histogram and tone mapping) whereas those characters can be easily recognized after biological enhancement inspired by photoreceptor model.

The enhanced frames resulted from tone mapping might seem to have better quality from regular viewer perspective. However, enhancing useful information is the main priority in video surveillance applications not the pleasant view experience (Comparison of conventional method with photoreceptor is described further in IV.8.6).
The main contrasting feature of this method is that it is uses temporal information to spatially enhance videos. One might question the comparison approach in which this method is only compared with spatial enhancement methods rather than temporal methods. Firstly, most of the commonly used contrast enhancements methods use only spatial information [147]. Secondly, temporal enhancement methods proposed in the literature are not specifically used for contrast enhancement [148]. It is therefore non-trivial to compare this method to conventional contrast enhancement. To achieve such a comparison, we compare the performance of an object classification system using no pre-enhancement, biological photoreceptor-based enhancement and tone mapping enhancement.

**IV.4 BRIEF DISCUSSION OF TIME COMPLEXITY**

In this section we briefly study the time complexity of the conventional and biological methods described in this chapter.

**Definition:** If $N$ is the set of natural numbers and $X^{mn}$ is a subset of points $(x, y) \in N^2$, therefore $m$ and $n$ denote the dimensions of $X^{mn}$ (i.e. $0 \leq x < m$ and $0 \leq y < n$).

For a point $(x, y) \in X^{mn}$, $t = I(x, y)$ is the level or intensity of the point in $I$. A mapping $I$ (from $X^{mn}$ to $N_L$) is called a monochromatic image. In this mapping $N_L = \{0, ..., L - 1\}$ and $L$ is typically 256 in most application.

As described previously, histogram equalisation works in three main phases:

1. Histogram computation
2. Cumulative density function
3. Contrast enhancement stage

Single scan throughout the image is required and hence it has $O(m \times n)$ time complexity. The second phase is dependent on the dimension of used histogram. Assuming that 1D histogram is used, the probability function can be computed in $O(L)$.

All pixels are processed in the third phase and therefore $O(m \times n)$ is the time complexity for this phase. One remark is that histogram equalization has $O(\max(L, m \times n))$ time complexity.

In tone mapping enhanced pixel value is calculated using Equation IV-2. All pixels will be processed in this method and its time complexity is simply $O(m \times n)$. Similarly all pixels should be scanned in biological contrast enhancement approach. Tone mapping and the biomimetic method however go through less complex phases and therefore take less time to generate an output frame. Histogram equalisation also needs some memory to retain pixel intensities and the number of occurrence in each image. Saving and retrieving these data will also introduce some delay to the contrast enhancement.
process. In summary, the bio-inspired photoreceptor method has no worse complexity than tone mapping, but offers superior performance (in the context of pattern recognition).

Table IV-1 time complexity of the conventional and biological methods

<table>
<thead>
<tr>
<th></th>
<th>Histogram Equalisation</th>
<th>Tone Mapping</th>
<th>Photoreceptor</th>
</tr>
</thead>
<tbody>
<tr>
<td>$O(\text{Max}(L,m \times n))$</td>
<td>$O(m \times n)$</td>
<td>$O(m \times n)$</td>
<td></td>
</tr>
</tbody>
</table>

In tone mapping and the biomimetic method, enhanced pixels can be independently obtained from original pixels without loading all other pixels. However, all pixels are firstly loaded into register in histogram equalization implementation in order to perform the algorithm. Hence, histogram equalization is not a suitable candidate for parallel processing that has been recently introduced in image processing.

Graphics Processing Units (GPUs) are designed to process and produce pixels for display on the screen. They are designed to solve very specific highly parallelizable problems. Streaming processors (SPs) are able to perform the same instructions on multiple data simultaneously. GPUs have become very popular in recent years due to their high performance parallel architecture and that is why biomimetic vision algorithms can be easily implemented in GPUs, which can significantly increase the algorithm performance and hence reduce processing time.

There are benefits in using a GPU as the main processing unit including:

- Acceleration of parallel workloads
- More efficient load-balance across system resources through the use of compute APIs designed for concurrency
- Freeing up CPU resource by offloading to GPU

GPUs are only able to process independent vertices and fragments; however they can handle many of them in parallel. This is very effective when we want to process many vertices or fragments in the same way. The photoreceptor model algorithm is well suited to parallel processing as luminance pixels are individually processed with no dependency on neighbouring pixels. In fact its temporal characteristics make it a good candidate for parallel processing. It is our view that employing GPU can significantly improve this bio-inspired enhancement in comparison with conventional methods [149].
CHAPTER IV – PHOTORECEPTOR AND COMPLEX LIGHTING

IV.5 PHOTORECEPTOR MODEL DESCRIPTION

The nonlinear features of the biological visual system which are evident in many components including the photoreceptor, can play an important role in the coding of visual information [150]. The light adaptation mechanism can effectively change the luminance range in which the visual system operates. Biochemical processes in the photoreceptor cells can competently perform the abovementioned adaptation tasks [151, 152].

Large illumination variation, due to change of season or day and night cycle, are common in the natural environment. This variation can also be generated in a shorter time scale; moving through the different parts of landscape, such as going from areas exposed to sunlight to the shade of a large umbrella. Shifting the eye gaze from the lit to the shaded section of the scene can be another example of short-term illumination variation.

The photoreceptor normally deals with at least two- to three-orders of magnitude in a fairly short time [153].

Many species are able to adjust their gain to light level changes in different environments. We describe and evaluate gain control in a photoreceptor model of that of blowflies. Such models are able to deal with a wide range of intensities. The main reason behind analysis of the photoreceptor model is its usefulness as a pre-processing module for studying visual processing in higher parts of the biological vision system. Adequate pre-processing, including gain control, is strictly required for higher visual processing stages. By studying the photoreceptor model (see Figure IV-2), we would be able to understand the properties of early-stage visual processing and how it reacts to natural stimuli.

Although previous linear and quasilinear systems proposed in the information theory literature have made significant progress [154-156], a good nonlinear model could be applicable in a more general range of applications.

Relatively simple models have been proposed to mimic light adaptation in blowfly phototransduction. Another advantage of such simple models is that fewer parameters need to be adjusted and can be applied in real time vision applications. Early-stage vision components such as the photoreceptor will serve as the pre-processing module. The natural series of intensities are normally used to mimic the intensity statistics that the photoreceptor will encounter in real environments [153]. The model that is employed in this work has been accurately inspired by an expected coherence function obtained by looking at response repeatability. This model is a cascade of several dynamic and static nonlinearities and is based on a parametric model proposed by van Hateren and Snippe [132] and modified by Mah et al. [157]. This stage dynamically controls several parameters of pixel intensities. Gain control, gamma correction, and non-linear saturating functions are the main control parameters in this model. The luminance dynamic range is compressed by gain increase in dark areas and gain reduction in lit areas. This light adaptation feature of the photoreceptor promotes the
detection of small moving targets in clutter. The initial parameters in this model are obtained from Mah et al. model [157]. Note that some of these parameters are modified later in order to improve the performance.

The photoreceptor algorithm is able to compress the luminance high dynamic range to low dynamic range by which not only the dynamic range is reduced but also the features of objects are enhanced in the process. Although this model does not use any spatial filtering, it can provide a form of high pass filtering in temporal variations to enhance moving objects. The advantage over conventional spatial filtering is that it will not generate noise by amplifying high frequency details. Also, by utilizing temporal information (i.e. using frame history) it can also reduce the noise and reintroduce features that may be missing in the frame.

The visible effect after applying the photoreceptor model is that moving objects will be enhanced while temporally stationary objects such as background regions will gradually fade (i.e. although there will always be a difference in their steady-state value).

Movement of the camera or objects within the scenes can highlight object features and the stationary scene will become noticeably duller over time (see Figure IV-3).

Figure IV-2 Photoreceptor Model diagram
In this model, each pixel is processed in isolation from others to achieve per pixel luminance control. Luminance variation can be classified into short term and long term [132]. The insect photoreceptor response has an initial fast transient (short term adaptation) and it will continue with slow decay for quite a long time (long term adaptation). Each of these variations gives rise to different aspects of the photoreceptor model. This model represents one photoreceptor cell process which is analogous to a single pixel in any image and thus to process an image many photoreceptor cells are required. This is very similar to the structure of the human eye in which the high dynamic range of the natural environment data is compressed by the non-linear adaptive model to view the surrounding area with more structural detail.

The model involves four stages of processing, three of which involve first order low pass filtering (the processing photoreceptor stages are demonstrated in Figure IV-4. Stage two plays an important role in this application. This in effect highlights motion due to luminance changes. This effect is produced via divisive feedback and involves a reasonably small time constant. As shown in Figure IV-7, time constants in this stage can considerably influence the step response. Since the photoreceptor is used to highlight structural details of moving objects, short-term adaptation needs to take place rapidly and thus a small time constant needs to be set for this stage. Luminance slow adaptation is modelled in stage three. Because of this, a larger time constant is used. This has a lesser impact on the complete model output, and so selection of the time constant is less critical.

*Figure IV-3 Photoreceptor Experiment (frame (1): no motion; frame (2): moving playing card)*

The photoreceptor makes use of the temporal information rather than spatial information and hence object features are only enhanced when there is motion in the scene.
The photoreceptor model stages are described in more details as follows.

**IV.5.1 Gain Control (Stage 1)**

In this stage the gain is varied based on light intensities. In other words, the gain increases when light intensity is lower and decreases in areas with higher light intensities. Another feature of this stage is its response speed. As it is shown in Figure IV-6, response of the signal is slower in darker areas and faster in lighter areas.

**IV.5.2 Divisive feedback temporal filtering (Stage 2 and Stage 3)**

Stage 2 features non-linear divisive feedback with low pass filter. This stage provides the short-term adaptation of the photoreceptor response. This response occurs over a very short time (i.e. a few milliseconds). This stage helps the system to adapt immediately when a cell is exposed to a rapid light intensity variation and can produce a logarithmic compression of input (see Figure IV-7). The output from the second stage is very similar to a square root functions in the steady state.

---

**Initialise**  
`Input_Stage1_LPF_PrevState`, `Input_Stage2_LPF_PrevState`, `Input_Stage3_LPF_PrevState`

**Stage1_LPF_PrevState**  
`= Input_Stage1_LPF_PrevState + Stage1_LPF_PrevState`

**Stage1_Output**  
`= (Dark_Gain \times (1 - \frac{Stage1_LPF_PrevState}{Input_Stage1_LPF_PrevState + Mid_Dark_Gain})) \times Input`

**Stage2_LPF_PrevState**  
`= Input_Stage2_LPF_PrevState + Stage2_LPF_PrevState`

**Divisor1**  
`= Stage2_LPF_PrevState`

**Stage2_Output**  
`= \frac{Stage1_Output}{Divisor1}`

**Stage3_LPF_PrevState**  
`= Input_Stage3_LPF_PrevState + Stage3_LPF_PrevState`

**Divisor2**  
`= k \times \exp\{Stage3_LPF_PrevState\}`

**Stage3_Output**  
`= \frac{Stage2_Output}{Divisor2}`

**Photoreceptor Output**  
`= \frac{Stage3_Output + Naka_Rushton_const}{Stage3_Output + Naka_Rushton_const}`

---

**Figure IV-4** this figure shows the processing pathway for a simplified photoreceptor model

Please note that time constants (i.e. Time_Const) in this algorithm are obtained from Brinkworth model [10]

`Input-stageX_LPF_PrevState` Initial values are set to “input frame” in first iteration

`PrevState` = Previous State, `LPF` = low pass filter
**Figure IV-5** coefficients impacts on Photoreceptor Model Stage Step Response ($T_2=$ Time constant (ms) that is used for stage 2 of photoreceptor)

**Figure IV-6** Stage 1 response in darker and lighter areas

This figure shows that if the input signal strength is higher (lighter areas) the response time is quick whereas it would take longer to respond if the input signal strength is low

(Note: This graph is an approximate stage response and demonstrates the functionality of this stage)
There is an exponentially weighted divisive feedback loop in the third stage and its low pass filter response has a slower response due to its long time constant. This stage provides the long-term adaptation to light condition variations over a longer time period (i.e. several seconds) and it shifts the operating range of the model.

**IV.5.3 Naka-Rushton transformation (Stage 4)**

The Naka-Rushton transformation is the final stage in the photoreceptor model and its response is determined by each pixel’s luminance [158].

In this stage a constant is added to the input and the result is used in a divisive feed forward operation. As it is shown in Figure IV-8, darker areas are further amplified in this final stage.

The output is more non-linear for lower luminance in the surrounding area and this can enhance the data at low luminance levels as a function of the immediate neighbourhood of each pixel.

From a biological perspective, weighted inputs to a neuron are interpreted as stimuli intensities. When the neuron is activated there is a spike rate and hence there is a gradual increase in firing until the maximum rate is reached.
IV.6 Temporal Filtering (Step Response)

If the video is considered as frames stacked in the three dimensions of x-y-time, we are able to temporally filter signals in the x-t or y-t planes. Such filters enhance object features in motion. In other words, stationary object features, in the x-y-t domain, form lines parallel to the t-axis, whereas moving features will trace a trajectory over time.

One of the most common filters used in insect vision models is the temporal first order low-pass filter. Temporal low pass filter parameters in the photoreceptor play an important role in this application.

Temporal filtering can be employed by implementing the following function:

\[ y(n) = y(n-1) + \beta (x(n) - y(n-1)) \]  \hspace{1cm} (IV-6)

Where \( \beta = \frac{\Delta T}{\tau + \Delta T} \), \( \Delta T \) is the time interval and \( \tau \) is the time constant. \( \beta \) has to be tuned properly to just enhance shape contours rather than introducing distortion from the previous frames. Choosing improper parameters can introduce extra edges or a visible ghosting effect in the enhanced frames. Three low pass filters have been used in this photoreceptor model and \( \beta \Delta T \) has to be tuned properly for different purposes. Derived from insect vision models, they will not be suitable for all applications.
For example, the step response of the photoreceptor with different coefficients (2\textsuperscript{nd} stage) is previously shown in Figure IV-5. By increasing the filter coefficient its response becomes slower and more of the frame history will appear on the photoreceptor output frame.

\section*{IV.7 Photoreceptor Implementation}

In the work presented in this chapter, the application of only the photoreceptor model is considered. To the best of our knowledge, this type of non-linear temporal processing has never been employed for shape recognition tasks.

For compatibility with conventional spatial-only shape recognition, the resulting perimeter enhancement is added to the original image data. The implementation algorithm is shown in Figure IV-9. In fact, the implementation of shape-recognition is based on only the object perimeter.

As previously explained, this model is applied to each individual pixel in any frame and thus to process an image many photoreceptor cells are required. This is very similar to the structure of the human eye in which the high dynamic range of the natural environment data is compressed by non-linear adaptation to view the surrounding area with more structural details.

In order to be effective for shape recognition applications, short term adaptation needs to take place rapidly and thus a small time constant needs to be set for this stage. Using large time constants in this stage can lead to identification errors.

Luminance slow adaptation is modelled in stage three. Because of this, a larger time constant is used compared to the fast adaptive nature of the previous stage. This has a lesser impact on the complete model output, and so selection of the time constant is less critical.

Three low pass filters have been used in this Photoreceptor model and $\Delta T$ has to be tuned properly to

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{flowchart.png}
\caption{This flowchart demonstrate how our Bio-inspired Video Enhancement method is employed to temporally enhance moving object features.}
\end{figure}

\begin{itemize}
\item In initial experiments it was noticed that in photoreceptor model stationary objects will gradually become duller when there is no motion in the scene therefore the original video is added to retain stationary object features. Gray scale mapping is also used to make it compatible with monitor display
\end{itemize}
improve the detection rate. Although parameters have been found in insect vision systems, these will not be suitable for this application as they can introduce large distortion from the previous frame which may lead to performance degradation. Shape contours from previous frames can have a negative impact on current frames which can appear as extra edges or a ghosting effect in the enhanced frame.

By using this biologically inspired method, shape contours become pronounced in different lighting conditions. In our experiments, $\Delta T_1 = 5, \Delta T_2 = 10\text{ ms}$ gave the best result among the considered candidate values. Initial parameters in this experiment are those given in the Brinkworth and O’Carroll Model [10]. In this model, $\Delta T_2$ has a significant impact on the sensitivity of enhancement method. Therefore, it has to be tuned properly. Looking at Figure IV-6, $\Delta T_2 = 10$ has a fairly sharp edge around step responses and that is why is selected as the time constant in this experiment. On the other hand, varying $\Delta T_1$ will not change the sensitivity of the system and $\Delta T_1 = 5\text{ ms}$ can be a reasonable time constant for this experiment.

Although some ghosting effect might be introduced in the obtained frames, such an effect does not impair shape recognition functionality. Recognition performance is the main focus here and shape appearance is a secondary issue.

IV.8 CARD RECOGNITION EXPERIMENT

IV.8.1 Automatic object recognition

Automatic object recognition system design has received considerable attention in the computer vision research community. Shape recognition is a special case and much progress has been made in developing robust shape descriptors to automatically identify different shapes from different viewpoints in still images or videos for various applications. This field of research is mostly motivated by industrial machine learning object identification systems, robotics and even automated surveillance systems [159-163]. The relative ease of obtaining moving object features from a distance makes these object identification systems very attractive tools. However, many of these automatic shape recognition systems perform poorly when applied in an uncontrolled environment. In other words, different illumination conditions, video quality, etc., can have a large effect on system performance.

Original video frames captured by camera sensors are normally compressed due to system bandwidth and storage constraints. Some artefacts introduced in compressed video sequences affect the video content, particularly the structural details of objects. Intra-frame variations due to changes in lighting conditions and sensor quality can also cause identification errors [164]. Most shape-recognition
methods are sensitive to variations in background shading and object illumination. Noise can also have a dramatic impact on recognition performance [165].

The way image sequences are encoded by technological systems, that is video, is fundamentally tied to the way in which the human eye and brain interpret images and motion. This includes such aspects as resolution, colour, dynamic range, frame rates and spatial and temporal compression techniques. On the contrary, object identification algorithms are commonly based on single image analysis, such as the extraction of a single video frame from a sequence. This mismatch of, in particular, temporal processing paradigms means that most object analysis algorithms are not well suited to the data with which they are presented. In order to bridge this gap we investigate the temporal preconditioning of video data through a biologically-inspired vision model, based on multi-stage processing analogous to the photoreceptors in insect vision systems. As opposed to other object identification algorithms which are based on single image analysis, in this work it is assumed that video is presented as a sequence of discrete still frames. In fact, the biological eye and brain interpret the sequence of frames as continuous motion and exploit motion in object identification [166-168].

As previously mentioned, frames are stacked in three dimensions of x-y-time form in video (Figure IV-10) and under these conditions feature extraction can be enhanced, thereby improving the accuracy of recognition systems.

This approach can lead to improved object identification through the enhancement of object perimeters and the amelioration of lighting and compression artefacts such as shadows and blockiness. In many shape recognition applications, the detection and identification of moving objects is of interest. In the x-y-t domain, stationary object features form lines parallel to the t-axis, whereas moving features will trace a trajectory over time.

In biological vision systems, such “temporal edges” are enhanced, thereby improving the visibility of moving objects compared to a stationary background. This approach can reduce the illumination variance for object detection and recognition systems. A similar approach is to use common edge detection filters in the x-t or y-t planes [169]. Such filters are intended to amplify object features in

![Figure IV-10 three dimensional spatiotemporal structure of video](image)

Single frames are run in succession to produce what appears to be a seamless piece of video. Individual frames contain spatial information however when they are stacked as above temporal information is also added to spatial information. Therefore, we can consider video as a three dimensional data (i.e. x-y-t)
In this work, we exploit temporal pre-processing which is inspired by photoreceptor behaviour to automatically enhance moving object structural details, leading to better shape recognition under diverse circumstances. Our approach is as follows:

1- First stage of insect vision system algorithm (i.e. photoreceptor) is implemented to pre-process video frames, specifically to enhance temporal edges for the purpose of performance improvement in shape recognition systems. Structural description is the main feature used in the tested system.

2- The video frame is assessed, with and without the employed temporal processing, using coefficients derived through a wavelet transform.

3- The proposed method is evaluated by implementation of a simple yet accurate original playing card suit recognition algorithm in order to compare the performance before and after temporal insect vision model implementation. Note that this recognition system is specifically designed for this application.

The experiment is described in three sections. We present and evaluate the accuracy of the shape recognition algorithm in the first section. The discrete wavelet transform is used to assess the structural shape details gained after using the proposed enhancement approach for the developed playing card classifier as explained in the second section and finally card suit recognition experiments are presented and discussed.

**IV.8.2 Shape classification algorithm**

We consider the practical problem of identification of the suit of conventional playing cards (see Figure IV-11) with obvious direct application in casinos, for example, and illustrative of a wider class of applications. We consider an algorithm which incorporates object shape features in a non-ideal environment. Many algorithms have been proposed for object recognition using such techniques as fuzzy, neural and feature-based identification systems [170-175]. Video complicates matters because any isolated frame is a two-dimensional projection of a three-dimensional object which is dependent on the angle of observation.
A robust recognition system has to show its invariability to the shape deformation which is caused by such arbitrary camera viewpoint.

We can model scaling, shearing, and translation of the object shape by using an affine transformation as follows:

\[
\begin{pmatrix}
    x'_a \\
    y'_a
\end{pmatrix} =
\begin{pmatrix}
    c_{11} & c_{12} \\
    c_{21} & c_{22}
\end{pmatrix}
\begin{pmatrix}
    x_a \\
    y_a
\end{pmatrix} +
\begin{pmatrix}
    b_x \\
    b_y
\end{pmatrix}
\]  \hspace{1cm} (IV-7)

Here, \((x_a, y_a)\) and \((x'_a, y'_a)\) are the coordinates of corresponding pixels on original and transformed shape respectively. Matrix \(C\) denotes an affine matrix that defines the rotation, scale, and degree of skew, while vector \(b\) represents a spatial shift. Similarity between different shapes can be measured by their features. Geometric features such as eccentricity, circularity ratio, elliptic variance, rectangularity, convexity, and solidity, etc., have been very popular to describe shapes. Using these geometric features, we would be able to discriminate shapes even from different points of view and distances from camera. For our purposes, the identification of the four suit types (heart, club, diamond, and spade) is a relatively small set and it is sufficient to use only a few simple descriptors for a relatively accurate classifier.
Our detector must be largely invariant to affine transformation. One widely used reliable metric used is solidity, which describes the extent to which the shape is concave or convex \([176]\), defined by:

\[
\text{Solidity} = \frac{A_s}{|H|}
\]  

(IV-8)

where \(A_s\) is the area of the shape region and \(|H|\) is the convex hull area of the shape. The solidity of a convex shape is always 1. While the diamond suit is purely convex, a heart shape is so close to being purely convex that under non-ideal viewing conditions these two suits can be easily confused. To overcome this issue, corners are mapped onto a normalized polar coordinate system centered at the object centroid to distinguish the shapes more effectively in a two-stage identification process \([177]\).

The flowchart of the algorithm is demonstrated in Figure IV-12. To test the efficacy of the proposed shape recognition algorithm, ideal noise-free video test sequences are synthesised (see Figure IV-13).

Four videos containing 2430 frames (30 frame/sec) were produced for each card suit (club, heart, spade and diamond) and all card suits underwent the same affine transformation (e.g. translation, rotation, scaling or shear, and in combination). We simply used affine transformation to generate the synthetic video sequence on which different measures of classification can be applied.

---

**Figure IV-12 Card Suit Classifier Algorithm Flowchart**

This algorithm is specifically tailored for this application (i.e. card suit recognition).

Threshold (Thr) Values in our Experiment: Thr0 = 0.81

Club (Thr1: 0.77, Thr2: 0.83), Spade (Thr1: 0.85, Thr2: 0.89),

Heart (Thr1: 0.90, Thr2: 0.95);
The algorithm classifies into one of five categories: spade, diamond, club, heart, and unknown. Under ideal conditions, the classifier achieved 100% correct recognition. But no shape classification system would have the same performance under non-ideal conditions. Most algorithms in the literature however are not tested in non-ideal complex-lighting conditions.

**IV.8.3 Polar coordinate implementation**

In a polar coordinate system each point is determined by a distance from a fixed point and its angle from a fixed direction. The fixed point, also known as pole, can be either origin of Cartesian system or another arbitrary point. The ray from the fixed point in the fixed direction is the polar axis and the distance from this pole is radius and the angle is polar angle.

Useful properties can be found in polar shape descriptors. Conversion from Cartesian to Polar coordinates will be advantageous in shape recognition algorithms, which can make them invariant to translation and scaling. Since the new coordinate is obtained by finding the distance from particular fixed point, translation of the object has no impact on polar shape descriptors. If we normalize the Euclidean distance according to largest distance value, such descriptors are invariant to scaling.

*Figure IV-13 Generated Synthetic Video frames with different affine transformation

**Synthetic is a good tool to analyze the algorithm performance in ideal conditions**

**In other words, it can simulate the ideal video in which no noise and artefact has been introduced**
The cartesian coordinate \((x, y)\) can be converted to polar coordinates \((r, \theta)\) by using below functions with \(r \geq 0\) and \(\theta \in (-\pi, \pi]\)

\[
r = \sqrt{x^2 + y^2} \quad \text{(IV-9)}
\]

\[
\theta = \arctan \left( \frac{y}{x} \right) \quad \text{(IV-10)}
\]

The two polar coordinates also can be converted to two Cartesian coordinates by using below functions:

\[
x = r \cos \theta \quad \text{(IV-11)}
\]

\[
y = r \sin \theta \quad \text{(IV-12)}
\]

The centroid is defined as the origin of the transform and it is obtained by:

\[
C = (C_x, C_y) = \left( \frac{1}{n} \sum_{i=1}^{n} x_i, \frac{1}{n} \sum_{i=1}^{n} y_i \right). \quad \text{(IV-13)}
\]

The centroid is the centre of gravity and its position will be fixed in relation to the shape. The number of points in the shape is given by \(n\). The Euclidean distance will be found between centroid \(C = (C_x, C_y)\) and boundary points \((x_i, y_i)\) and the resultant values will be put into two vectors \((\Theta^i\) for angles (in degrees) and \(P^i\) for radii). These boundary points are distributed on shape contour uniformly.

\[
\rho_i = \sqrt{(x_i - C_x)^2 + (y_i - C_y)^2}, \quad \text{(IV-14)}
\]

\[
\Theta_i = \arctan 2 \left( \frac{y_i - C_y}{x_i - C_x} \right). \quad \text{(IV-15)}
\]

Note that nearest integers of \(\Theta_i\) will be computed by:

\[
\Theta_i = \left\lfloor \Theta_i \right\rfloor \quad \text{if } \left| \Theta_i - \left\lfloor \Theta_i \right\rfloor \right| < 0.5
\]

\[
\Theta_i = \left\lceil \Theta_i \right\rceil \quad \text{if } \left| \Theta_i - \left\lceil \Theta_i \right\rceil \right| \geq 0.5 \quad \text{(IV-16)}
\]

In the resulting image the angular coordinates are placed on horizontal axis and the Euclidean distance coordinates are plotted vertically.

After polar projection of the boundary of each shape, we are able to find shape corners by finding local extrema as presented in Figure IV-14. We devised a polar descriptor (PD) in which the first two maxima, showing diamond corners, are found and their distance will be computed in Cartesian coordinates. Afterwards, this value is divided by half the perimeter.
\[ P_D = \frac{\text{Corner Dist}}{P/2}, P = \text{perimeter} \quad \text{(IV-17)} \]

Note that this projection is only useful for efficiently separating diamonds from the other three shapes and normalizing the coordinate system also makes this metric invariant to affine transformation.

**IV.8.4 Card segmentation process in real world**

Although synthetic video is a useful tool to test the function of the devised algorithm under desired conditions, it cannot clearly show algorithm effectiveness in real world scenarios. Therefore, a real world experiment is also presented here. Non-ideal conditions are generated by using a high dynamic range (12 bits per pixel) video camera when lighting is not fully controlled. Most light beams are emitted on cards from behind the camera and some shadows are present in this video.

A four-minute standard definition video (720x480 pixels at 25 frames per second) was made under realistic lighting conditions with arbitrary movement of four cards, one card at a time. One minute each was allocated to each card, each in a different suit, or approximately 1500 frames per card. In order to recognise card suits, cards would have to be segmented from background and in the second stage card suit shapes will be detected from segmented cards.

Recognizing playing cards under real conditions is much more complicated and involves segmentation, edge detection and other spatial image processing techniques. The accuracy of such a system is very dependent on lighting and noise circumstances. There have been only a few card

*Figure IV-14 Boundary Signature in Polar Coordinate*
recognition systems proposed in the literature in which non-ideal lighting and noise conditions were chosen as their laboratory setting [178, 179]. However this is the case in most scenarios especially if the system will be implemented in areas with even more complicated lighting conditions. The first task was to separate the card from background and to do this we used a simple background subtraction with adaptive thresholding (see Appendix A for more details) in the luminance plane. Since the background was darker compared with the card faces this was not very challenging and the major problem was shadows on the cards. Object segmentation processing is then employed to extract the card from the binary image. A median filter is also used to reduce the noise in the binary image. A binary closing operation is also used to fill the object within the interior region of the object edge contour. The output of this stage was a binary image with cards as white and background as black.

After finding the location of the card, a mask is used to extract the card data. Finally, an edge-detecting filter is used to extract playing card shapes for the playing card suit recognition algorithm. In fact two consecutive segmentation algorithms were executed to efficiently extract suit shapes from playing cards. The summary of feature extraction, and card segmentation, is shown in Figure IV-15.

**IV.8.5 Spatial frequency Analysis**

Spatial frequency analysis is a useful tool for subjective evaluation of object enhancement using pre-processing. In fact, frequency domain techniques could be used to perform the pre-processing stage of enhancement of structural details, particularly contours, through manipulation of high frequency components in the image such as by convolution with an appropriate kernel. However such implementation requires the selection of a particular edge detection filter, which is application-specific. It is therefore difficult to extract all edges reliability under a variety of realistic conditions.

The wavelet transform is a reliable tool for the extraction of details as well as approximations of images, such as in the JPEG2000 image representation standard [180].

A two-dimensional two-level Daubechies wavelet transform enable the decomposition of an image into several space-frequency sub-bands.
**IV.8.5.1 Why use Discrete Wavelet?**

The wavelet transform is composed of two parts, a low-pass filter (scaling) and a high-pass filter (wavelet). The wavelet transform can simultaneously examine both the position and the frequency of the signal. The wavelet transform has been widely used in the field of signal processing, especially image compression. Our interest in the study of the wavelet transform centres on its usage in image analysis after employing the photoreceptor model. In fact the progressive reconstruction feature of this waveform makes it a suitable candidate to see how spatiotemporal processing impacts on different parts of the image in the frequency domain. In other words, this transform can precisely identify locations where activities occur in an image which sinusoidal transforms (e.g. DCT transform) will not be able to do. Therefore when we can distinguish areas of intense activity from flat regions and investigate how they are encoded in compression algorithms.

**IV.8.5.2 Discrete Wavelet Transform Implementation**

The discrete wavelet transform is used when the signal and scaling functions are discrete in space.

The DWT of a sequence comprised of two series expansion. The first component corresponds to the approximation and the other corresponds to the details of the sequence. The formal definition of DWT of an N-point sequence \( x[n], 0 \leq n \leq N-1 \) [181] is given by

\[
DWT\{ f(t) \} = W_\phi(j_0, k) + W_\psi(j, k)
\]

Where

---

![Figure IV-15 Card Segmentation Algorithm](image)

*Figure IV-15 Card Segmentation Algorithm*

This flowchart demonstrates the main components of card segmentation algorithm.

The size and sensitivity of the filters used in this algorithm (e.g. morphological filters) are based on the object size (i.e. playing cards, playing suit shapes).

In this experiment: Median filter (with the size of 12x12) was firstly implemented and then diamond-shape morphological filters (i.e. dilation (size of 8x8) and erosion (size of 4x4)) were applied to segment the card from the background.
CHAPTER IV – PHOTORECEPTOR AND COMPLEX LIGHTING

\[ W_\phi (j_0, k) = \frac{1}{\sqrt{N}} \sum_{n}^{N-1} y[n] \phi_{j_0,k}[n] \]  \hspace{1cm} (IV-19)

\[ W_\psi (j, k) = \frac{1}{\sqrt{N}} \sum_{n}^{N-1} y[n] \psi_{j,k}[n], \quad j \geq j_0 \]  \hspace{1cm} (IV-20)

The sequence \( y[n], 0 \leq n \leq N - 1 \) can be recovered from the DWT coefficients \( W_\phi \) and \( W_\psi \) as given by (see also Figure IV-16):

\[ y[n] = \frac{1}{\sqrt{N}} \sum_{k} W_\phi (j_0, k) \phi_{j_0,k}[n] + \frac{1}{\sqrt{N}} \sum_{j=j_0}^{\infty} \sum_{k} W_\psi (j, k) \psi_{j,k}[n] \]  \hspace{1cm} (IV-21)

The scale parameter in the second summation of equation (IV-22) has an infinite number of terms. But in practice the upper limit for the scale parameter is usually fixed at some value say \( J \). The starting scale value \( j_0 \) is usually set to zero and corresponds to original signal. Thus, the DWT coefficients for \( x[n], 0 \leq n \leq N - 1 \) are computed for \( j = 0,1,\ldots, J - 1 \) and \( k = 0,1,\ldots, 2^j \). Also \( N \) is typically a power of 2, of the form \( N = 2^J \).

In images a two-dimensional (2D) DWT is used. Since the 2D DWT is separable we can implement the 2D DWT in a row-column fashion. It is more efficient to implement the 2D DWT using 1D DWT and there are also algorithms for fast 1D implementation.

In (2D) DWT, the image is partitioned into four sub-bands: low-low (LL), low-high (LH), high-low (HL), and high-high (HH) as shown in Figure IV-17. The high frequency subbands contain

\[ \begin{array}{c}
\downarrow \downarrow \downarrow \downarrow \\
N \times N/2 \\
N \times N/2 \\
N \times N/2 \\
N \times N/2 \\
\end{array} \]

\[ \begin{array}{c}
h_0[-n] \\
h_0[-n] \\
h_0[-n] \\
h_0[-n] \\
\end{array} \]

\[ \begin{array}{c}
y_{LL}[m,n] \\
y_{LH}[m,n] \\
y_{HL}[m,n] \\
y_{HH}[m,n] \\
\end{array} \]

\[ \begin{array}{c}
h_0[-n] \\
h_0[-n] \\
h_0[-n] \\
h_0[-n] \\
\end{array} \]

\[ \begin{array}{c}
N/2 \times N/2 \\
N/2 \times N/2 \\
N/2 \times N/2 \\
N/2 \times N/2 \\
\end{array} \]

Figure IV-16 forward one-level 2D DWT via sub-band coding scheme
significant edge details. The idea to use DWT here is to combine the various DWT high-pass sub-bands in order to perform effective edge detection.

The decomposition procedure for such a redundant wavelet transform is different from the normal one in that the scaling of the wavelet is not achieved by sub-sampling the image in each step, but rather by an up-sampling of the filters. We construct wavelet sub-bands of the same size as the original image by up-sampling techniques (i.e. nearest-neighbour method is used in our experiment to approximate the missing pixel values). To take advantage of the multi-resolution property of the wavelet transform, we can superimpose each pair of corresponding sub-bands at adjacent resolution in order to compare image details. Lower edges, appearing in high frequency components, can also be filtered out in order to suppress noise. In fact, edge structures are present at each scale while noise decreases substantially along the scales, as shown in Figure IV-18.

The edge map $F$ is thus obtained by addition of the three high-frequency sub-bands $LH$, $HL$, and $HH$ together, with appropriate up-sampling.

$$F = \{ (LH1 + LH2) + (HL1 + HL2) + (HH1 + HH2) \}.$$  \hspace{2cm} (IV-23)

### IV.8.6 Statistical Analysis

The algorithms and tools used in our experiments for shape classification and analysis rely on a high number of parameters in order to work properly. In fact, good parameter optimization (i.e. so-called system calibration) can yield an efficient and stable system. In our experiments, the well-tuned parameters were obtained during the algorithm development stages which enabled us to make the designed system compatible with the application. That being said, the main purpose of this research is
not to propose a new shape recognition system. Shape recognition system is employed to demonstrate the effectiveness of biological enhancement. In other words, our aim was to show how biological vision system can aid us to improve our designed recognition system in non-ideal conditions. Therefore, we investigated how the designed system can perform under ideal conditions.

Having established perfect classification under ideal synthetic video conditions, the classifier was tested using real footage. Classification decision thresholds were set such that the false alarm rate was...
no more than 10% in order to evaluate the system with and without the photoreceptor model. Under these real conditions, the classification rate decreased to approximately 75% correct classification.

When the biologically-inspired vision pre-processing stage is inserted before the card segmentation algorithm, however, the probability of detection improves significantly, typically to 82%, as shown in Table IV-2 and Figure IV-19. This result was achieved, furthermore, without empirical optimisation of photoreceptor parameters.

In fact, the biological model needs to be calibrated for each recognition system to avoid negative impacts on used shape recognition performance. Further consideration of temporal filter parameters, which could have both positive and negative impacts on classification performance, is proposed for future work.

By way of performance comparison, a conventional global tone-mapping approach was used for spatial-only enhancement prior to classification of the same footage. Such pre-processing was not able to produce an improvement of statistical significance (see Figure IV-19 (95% confidence levels overlap)).

In summary, a novel adaptive biologically-inspired video pre-processing scheme has been proposed which enhances frames for later shape identification analysis under a wide variety of image conditions. We also developed a simple yet robust card suit recognition system to be able to show our proposed enhancement method impacts on shape recognition systems. The shape descriptor used for card suit classification was novel and its invariance and accuracy in various form of affine transformation were demonstrated.

Table IV-2 Recognition Rate Comparisons between Original and Enhance video frame (Mean and [95% Confidence Interval]) expressed as a percentage

<table>
<thead>
<tr>
<th>Enhancement Method</th>
<th>Clubs</th>
<th>Heart</th>
<th>Spade</th>
<th>Diamonds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>75.6</td>
<td>73.4</td>
<td>77.2</td>
<td>72.6</td>
</tr>
<tr>
<td></td>
<td>[73.3  77.7]</td>
<td>[71.0  75.6]</td>
<td>[74.9  79.3]</td>
<td>[70.3  74.8]</td>
</tr>
<tr>
<td>Conventional Methods</td>
<td>76.4</td>
<td>74.8</td>
<td>78.2</td>
<td>74.7</td>
</tr>
<tr>
<td>(tone mapping)</td>
<td>[74.1  78.6]</td>
<td>[72.2  77.3]</td>
<td>[76.5  79.8]</td>
<td>[73.0  76.4]</td>
</tr>
<tr>
<td>Photoreceptor</td>
<td>83.1</td>
<td>82.4</td>
<td>83.7</td>
<td>82.2</td>
</tr>
<tr>
<td></td>
<td>[81.1  85.0]</td>
<td>[80.4  84.3]</td>
<td>[81.7  85.5]</td>
<td>[80.2  84.1]</td>
</tr>
</tbody>
</table>
The biological model enhances contrast and edges to mitigate lighting variations. Since the photoreceptor model adaptively compresses intensity variation, bright and dark regions receive much the same attention for further processing. Therefore even under poor illuminations moving object contours are pronounced. Experimental results show that the proposed method improves the card suit recognition performance, and yet under the same conditions the use of a widely used conventional approach shows marginal improvement. This is to be expected to some extent, since the biological model incorporates history from previous frames and so bases the classification decision on a larger data set. The biological model does this without requiring training or parameter selection based on a-priori knowledge of lighting conditions. It should be noted however that enhancing object contours by using temporal edges cannot always help shape recognition systems and it might introduce some noises or unwanted edges that can even cause the system performance to deteriorate.

Figure IV-19 the confidence interval of each method is depicted for better view of their performance. It can be seen that conventional method (i.e. tone mapping) can slightly improve the recognition rate, but this improvement is not significant (i.e. there is an overlap between conventional and original card suits confidence interval).
Figure IV-20 Showing frames enhanced by photoreceptor model (above) versus unenhanced (original) frames

The card suit in frame 4 is incorrectly recognized as spade rather than club.

Temporal preconditioning helps to improve card recognition performance in difficult scenarios (card suit is correctly recognized in frame 2).

In fact, it would be challenging for the recognition system to recognize card suits when playing cards are far from the camera.
Chapter V  LMC AND NON-LINEAR SPATIO-TEMPORAL ENHANCEMENT

V.1 Problem Statement

Digital video is widely used as digital evidence. Videos are compressed in order to reduce the file size or save bandwidth in transmission. However, lossy compression can discard detail in video footage which might be of relevance in a forensic analysis of the video content, such as recognizing alphanumeric characters. This can cause incorrect identification of letters, numbers and other shapes such as faces. Visible artefacts caused by lossy compression raise questions of the validity and therefore admissibility of such evidence.

It is common for digital video to contain visible artefacts due to lossy compression, such as blurriness, blockiness, smearing and repetition of objects.

Blockiness for example is caused by the loss of high spatial frequency information. Translation of artefacts, due to motion-based compression, is another major source of distortion which can also affect object shapes and therefore object recognition, such as that of alphanumeric characters.

Whether or not there is subsequent editing and recompression of the video, the reliability of any digital video presented in court has been questioned because of the potential for content alteration during compression stages [108].

Video might be modified intentionally, innocently, or by the inadvertent introduction of compression artefacts. Specifically, it has been demonstrated that lossy compression alone can be the main source of information alteration [4].

Obtained video (or image) quality has to be carefully examined in order to make sure information alteration has not had significant impact on its fidelity.

Considerable effort has gone into the development of video and image quality optimization under lossy conditions. However, such optimization deals almost exclusively with the noticeable psycho-visual impact rather than the retention of detail that is the subject of forensic analysis [182]. In fact, the majority of the studies in this field rely on assessing the resulting compressed video based on objective and subjective quality metrics of visual impact.

The most widely used objective video quality metrics are mean squared error (MSE) and peak signal to noise ratio (PSNR) which play important roles in various video processing applications [183]. For subjective methods, non-expert individuals typically either visually compare the compressed video or image with the original, or provide an opinion based only on the compressed version without reference to the original. In both cases the purpose is to ensure a pleasant viewing experience for the human eye, rather than the retention of detail that might be of relevance in a forensic context. In this
work, we are concerned with after-the-fact analysis of compressed video, such as that captured by surveillance cameras. In addition to the introduction of compression artefacts, such automated or manual analysis is also affected by optical noise, lighting conditions, image resolution, and other factors.

The specific example of identifying alphanumeric characters in a video sequence has been investigated in this work. The main aim is to improve compressed video to ensure sufficient details for later analysis has been captured.

Many characters might be blended into the background in compressed videos. Colours can also change spatially and temporally. Another common case is when stationary texts jump around by a few pixels. Although the abovementioned distortions may seem unimportant, these can challenge recognition tasks. A typical example is given in Figure V-1. The difficulty with a subjective quality review is that even though the compressed view of the number plate appears to be readable at the original scale, it is clear in the enlargement that so much detail has been lost that the text is unreadable. In fact, forensic experts should be able to intuitively recognise characters and it is hard to recognise characters without looking at the uncompressed frame.

![Figure V-1 Confusion in Number Plate Recognition under JPEG Compression](image)

*This is the scenario in which artefacts introduced by video compression have fully distorted the object details which cannot be restored*
Our objective is therefore to introduce into the compression process a means of ensuring alphanumeric shape fidelity whilst retaining compatibility with the compression codec and constraints on memory or bandwidth.

In this work, we exploit a part of an insect vision system model to automatically enhance changes in space and time, thereby highlighting the object boundary points. Our novel pre-compression video enhancement method is inspired by the properties of the Large Monopolar Cell (LMC).

Our hypothesis is that more semantic information would be retained at comparable data rates, thereby improving recognition rates and integration of standard compression schemes. In this experiment, it is demonstrated that integration of biologically-inspired video enhancement can lead to recognition accuracy improvement. The effectiveness and limitations of the proposed approach are also thoroughly explained in this chapter. It is shown that such enhancement overcomes moderate video compression distortion effects such as JPEG and motion compensation artefacts.

The approach is as follows:

1. We firstly investigate the inter-frame and intra-frame compression impacts on alphanumeric characters. A synthetic video (ground truth) is used to explore the DCT block-based compression influences on alphanumeric characters.

2. An LMC-inspired vision system algorithm is then implemented to pre-process video frames, specifically to enhance moving object features, prior to lossy compression.

3. We perform cross-correlation of the Fourier descriptor of compressed video against letter templates and compare the correlation performance for compressed video with and without pre-processing enhancement. If the enhancement technique is effective, this will be reflected in improved correlation.

V.1.1 Challenges for surveillance applications

There has been significant progress in building license plate recognition systems and there is still a lot of work to be done in this area. In fact, a robust system that can work effectively under a variety of lighting conditions has not, to our knowledge, been devised. Such a system should be able to deal with number plate recognition with the low level of detail captured in low light conditions. Coupled with the artefacts introduced by video compression algorithms to alpha numeral characters, correct recognition becomes even more challenging.

Frames are captured and analysed without reference to other frames in such recognition systems. However, taking advantage of the temporal information of a video can, we hypothesise, improve the system performance. Some of the higher spatial frequency details lost in the compression process can be recovered and enhanced, thus potentially improving classification performance. Several different strategies can be applied to achieve this goal. One such approach is generating a high resolution image
by combining multiple sub-pixel shifted frames. However, this approach is computationally expensive and might not be applicable in real time applications including surveillance technology.

One approach, inspired by the insect vision system, is to only use previous frame information to enhance moving structural details within the current frame. Obviously there are many other open issues for future research. Future systems will likely to be equipped with faster processors and that will help designers to implement more complex and hence efficient systems. However, an effective system should be able to integrate with a variety of existing surveillance equipment.

Extracting an image from a low quality video sequence that can offer an acceptable level of detail can be very challenging at times especially if the duration of video is short and the object speed is fast. To deal with the illumination problem, a good enhancement method should be implemented to alleviate the influence of lighting and to make the object more visible.

Another major issue with previous proposed methods is that they have not focused on ambiguous characters that can be misinterpreted in the recognition stage [184]. However it is not the intention of this thesis to compare previously proposed character recognition systems.

Another challenge is to evaluate the performance of each method in each stage (i.e. license plate extraction, segmentation etc.) as there is no standard set of videos. Our research though has no intention to compare the automatic license plate recognition (ALPR) systems proposed in the literature. We are however investigating the impact of lighting conditions and video compression deterioration on alpha-numeral characters. A novel biologically inspired video enhancement method is then introduced that can potentially provide more information for ALPR systems and hence improve their performance. This is demonstrated by reporting improvements in correlation against alpha-numeral templates.

V.1.2 Conventional techniques

In most automatic license plate recognition systems a small set of still images are captured and analysed independently. Most ALPR systems comprise main three stages (see Figure V-2):

1. license plate extraction
2. character segmentation
3. character recognition process.

Segmentation and recognition of characters can become a difficult task if the video has been highly compressed or is captured under low lighting conditions. Therefore, different spatial enhancement methods might be applied to the extracted license plate. Different spatial filters can be used in order to enhance alpha numeral characters located on the license plate. The performance of some of these filters were compared in the literature [185], in which Sobel, Kirsch, Laplace, Canny, Rothwell and
SUSAN filters were compared to enhance character edges. The Canny edge detector performed best in this research.

![Figure V-2 Typical automatic license plate recognition system](image)

Note: spatial enhancement (filtering) might be also used in other stages of character recognition systems based on its model design

Although the main goal of this research is to exploit the spatiotemporal biological vision system model in current video surveillance technology, it is appropriate to compare our model with current spatial approaches used in ALPR systems.

### V.1.2.1 Canny Edge Detector

One of the fundamentals in spatial processing is finding edges in an image in which areas with strong intensity contrast (i.e. large variation of pixel intensity) are extracted. In this way some redundant data is filtered out while structural details of the image are preserved. Sobel, Laplace and Canny edge detection algorithms are popular methods in image processing of which the Canny edge detector is known as the most optimal edge detector [186].

Before implementing the Canny edge detection algorithm, the image needs to be smoothed in order to suppress existing noise. Gaussian filtering is normally used as a low pass filter due its simplicity. In this step, a suitable mask is calculated and low pass filtering is carried out using standard convolution methods. Convolution masks are usually smaller than the actual image. Large Gaussian masks are not very sensitive to image noise. In other words, more localization error is introduced if the Gaussian width is increased. They are also computationally expensive compared to small Gaussian masks.
After the smoothing stage, edge strength is found by taking the image gradient. Since there are edges in a variety of directions, four filters are used to detect vertical, horizontal and diagonal edges in the blurred image.

The first derivative (for horizontal and vertical direction) carried out in the edge detection algorithm returns values of gradients \( G_x \) and \( G_y \). Direction and gradient can be computed from these values as follows:

\[
G = \sqrt{G_x^2 + G_y^2} \quad (V-1)
\]

\[
\Theta = \arctan\left(\frac{G_y}{G_x}\right) \quad (V-2)
\]

\section*{V.2 LMC Model Introduction}

Laminar Monopolar Cells (LMCs) is very similar to bipolar cells in mammalian eyes \cite{125} and models designed for additional processing that is performed in second-order neurons in flies. This stage helps to discard the redundant details in both the space and time domain in an efficient way. Parameters in LMC filters are varied in both space and time depending on light levels. For instance, there is higher cut-off in areas of higher illumination. Finally a non-linear saturation transform is introduced that can improve the system performance by compressing higher and lower pixel intensities. Spatial high-pass filtering used in this model is defined as the response from the neighbouring pixels on the hexagonal pixel grid \cite{10}. Different types of processing (e.g. attenuation, inversion etc.) will be applied on such neighbouring pixels before they can be combined with the signal from the centre pixel \cite{187}.

Experiments show that inclusion of this stage after the photoreceptor could improve motion detection and velocity estimation (executed in the later Elementary Motion Detection (EMD) model) \cite{10}. Processing of the LMC with its spatial-temporal high pass filtering feature has been considered as one of the most important pre-processing elements of motion detection systems in the insect vision system (see Figure V-3). However, this stage can only be effective if it is combined with other stages such as the photoreceptor and it will not provide significant improvement if it is used by itself \cite{188}. Being beneficial to the motion processing stage, the LMC also effectively optimises information transmission in limited bandwidths.
Although a combination of spatial and temporal filtering has been used in insect vision models, they have different roles to enhance object features. Spatial filtering is performed on standalone frames to separate spatial frequency information. For instance, in spatial high pass filters (employed in this model), high spatial frequency information is retained within a frame while the low frequency information is discarded. The biological eye and brain interpret the sequence of frames as continuous motion.

The model includes both low-order spatial high-pass and temporal low-pass filters. Signal amplifiers are also included in this stage. The spatial filters are used to emphasize edges and areas of relative movement.

Figure V-3 Algorithm and Schematic of Implemented LMC Model (Note: High pass filter in step 3 is temporal)
V.2.1.1 Spatial Filtering in LMC
To implement less computationally expensive spatial high pass filtering in a biologically-inspired system model, the general procedure is to apply a low pass filter to the original frame and subtract this low frequency image from the original.

A Gaussian low pass filter is widely used in many computer vision algorithms for image structure enhancement at different scales.

An \( n \times n \) Gaussian filter (i.e. 5x5 in our experiments) is normally used in model implementations.

\[
f_H(x, y) = f(x, y) - f_L(x, y) \quad (V-3)
\]

Where \( f_L(x, y) \) is spatial low pass filtered image (Gaussian Blur) and \( f_H(x, y) \) is the resulting high pass filtered frame.

V.2.2 Post-Compression or Pre-Compression?
Since we had the option of using LMC as pre-compression or post-compression processing block, both were explored individually to find the more efficient system. We assumed that biological spatiotemporal filtering would not be able to distinguish between introduced types of noise. In fact, lossy video compression algorithms introduce a significant amount of distortion to raw video; therefore, post-processing could make it even more complicated as the distortion is unfavourably amplified. Our expectation was that pre-compression processing (compared with post-processing) would lead to better results (see Figure V-4). Our preliminary results validated our assumptions and no evidence of improvement was identified in using the biological LMC model as a post-processing block. 50 frames were randomly selected from synthetic video and the character “G” is chosen for this brief experiment (see Figure V-5). Auto-correlation is used to test if LMC works as a post-processing algorithm.
In our experiments, LMC is used as both post- and pre-processing video compression component in order to compare its effectiveness. Initial results however showed no improvement in using LMC as post-processing component.

Autocorrelation declined after introducing LMC as a post-processing stage. These results suggest that LMC cannot be used as post-processing video compression component.

Note: other five characters (i.e. C, H, N, O, and Q) auto-correlations are also analyzed in our experiment and the obtained result had similar trend to this one.
V.3 EXPERIMENT DESCRIPTION

Shape is considered one of the most important features that helps humans to recognise different characters according to a range of models [189]. In order to retrieve and recognise shapes, different shape representation approaches have been used. Among them, the Fourier descriptor technique is widely used. Moreover, the distortion effects caused by compression on character shapes can also be studied by comparing such descriptors.

The Fourier descriptor generally refers to the use of the Fourier transform to analyse a closed planar curve. The use of Fourier descriptors as a representation for closed curves was first proposed by Cosgriff [190]. Such shape signature receives much attention in the literature, specifically in shape identification applications [191-193]. The boundary can be uniquely represented by series of Fourier coefficients. In practice this series can be truncated to give a finite shape descriptor whilst still retaining sufficient descriptor power. The lower order of this series describes the macroscopic behaviour of the boundary curve. As demonstrated in Figure V-6, by keeping only a subset of the lowest frequency coefficients we can extract its shape signature [194]. This approach has been shown to be effective in similar applications [195, 196]. By using the Fourier descriptor here we are able to visualize how much of the lower order series of these coefficients has been lost during compression.

For the purpose of demonstration, the FE-Schrift font, specified for vehicle registration plates in Germany since 2000 [197], was used.

<table>
<thead>
<tr>
<th>Uncompressed</th>
<th>Compressed</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Characters</strong></td>
<td><strong>Characters</strong></td>
</tr>
<tr>
<td>A B C D E F</td>
<td>A B C D E F</td>
</tr>
<tr>
<td>G H I J K L</td>
<td>G H I J K L</td>
</tr>
<tr>
<td>M N O P Q R</td>
<td>M N O P Q R</td>
</tr>
<tr>
<td>S T U V W X</td>
<td>S T U V W X</td>
</tr>
<tr>
<td>Y Z 0 1 2 3</td>
<td>Y Z 0 1 2 3</td>
</tr>
<tr>
<td>4 5 6 7 8 9</td>
<td>4 5 6 7 8 9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Fourier Descriptors</strong></th>
<th><strong>Fourier Descriptors</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>A B C D E F</td>
<td>A B C D E F</td>
</tr>
<tr>
<td>G H I J K L</td>
<td>G H I J K L</td>
</tr>
<tr>
<td>M N O P Q R</td>
<td>M N O P Q R</td>
</tr>
<tr>
<td>S T U V W X</td>
<td>S T U V W X</td>
</tr>
<tr>
<td>Y Z 0 1 2 3</td>
<td>Y Z 0 1 2 3</td>
</tr>
<tr>
<td>4 5 6 7 8 9</td>
<td>4 5 6 7 8 9</td>
</tr>
</tbody>
</table>

*Figure V-6 Character Fourier Descriptors*

1st column: Uncompressed 2nd column: Compressed
1st row: Alpha-numeral Characters 2nd row: Fourier Descriptors (15 lowest-order coefficients)
This font was designed to be hard to manipulate, for example by adding or removing character features to change a licence number and is relevant in this application because the character differences should also be conducive to effective automatic identification. A set of 36 alphanumeric templates each of 34 by 34 pixels was constructed, with characters centred in each template (see Figure V-7). Above this resolution alphanumeric characters generally have sufficient details, even under extreme compression, that character recognition is not significantly affected. However as the resolution decreases JPEG compression artefacts can have a significant impact on their signature.

As a first step, cross-correlation between each of the 36 alphanumeric characters was considered. A high cross-correlation (approaching 1) indicates that two characters are very similar. This preliminary computation is used to highlight the character subsets most likely to be difficult to distinguish. As shown in Table V-1, it is clear that the most ambiguous character subsets are G and C; M, N and H; and 0, O and Q. There is also very high correlation between 0 and D; and 0 and G; and high correlation between, for example, Q and C; and O and C.
**CHAPTER V – LMC AND NON-LINEAR SPATIO TEMPORAL ENHANCEMENT**

Table V-1 Alpha-Numerical Characters correlation matrix (without compression)

Group 1 (Red): Very High Correlation e.g. (G and C)-(M, N and H) – (0, O and Q) ≥ 0.8

Group 2 (Orange): High Correlation e.g. (0 and D)-(0 and G)-(C and B) ≥ 0.8

Group 3 (Blue): Considerable Correlation e.g. (Q and C)-(O and C) ≥ 0.75

---

A synthetic test video sequence was developed (see Figure V-8) to measure the proposed video enhancement under ideal conditions. Each alphanumeric card moves horizontally while the size of the character changes from frame to frame, from 24x24 to 72x72 pixels. This represents, for example, the changes in the visibility of a number plate as a car approaches or moves away from the viewer but assumes a full frontal and unrotated view. The video sequence also incorporates smooth lighting (and therefore contrast) changes. The alphanumeric cards have stronger contrast (lighter background) where the background of the field of view is lightest; and lower contrast (darker background) where the field is darkest. There is a slight luminance shift between the field of view background and the card background to assist object detection. Each video contained 610 frames at 25 frames per second. The Canny edge detection algorithm [198] is used to find boundary points of the characters. This effective edge detector commonly performs better than others unless conditions are particularly suitable [199].
After obtaining character contours by the Canny edge detector, coordinates of pixels along the circumference are sorted clockwise around the centre and the coordinates are expressed as a vector of complex numbers. A Discrete Fourier Transform is applied to the vector and high frequency components are discarded. Inverting the process can result in the so-called Fourier descriptor of each character.

V.3.1 Fourier Descriptor:
The Fourier descriptor as its name suggest is used to describe the boundary of a shape in two-dimensional space by using Fourier transform.

After object segmentation and boundary extraction $N$ samples of a shape boundary on the xy-plane are taken. This procedure is done by travelling the boundary with constant rate and periodic coordinate measurements. The coordinates are then mapped onto the complex plane:

$$s(k) = x_k + iy_k \quad (V-4)$$

This representation will change its dimension from 2-D into 1-D and instead of having $N$ complex numbers we have $2N$ real numbers.

The Discrete Fourier Transform will be used to find Fourier descriptors of the boundary.

---

Figure V-8 Synthetically Developed Video (fr=frame number)
CHAPTER V – LMC AND NON-LINEAR SPATIO TEMPORAL ENHANCEMENT

\[
a(u) = \frac{1}{N} \sum_{k=0}^{N-1} s(k) e^{-\frac{i2\pi k u}{N}}, \quad u = 0, 1, 2, \ldots, N - 1 \quad (V-5)
\]

The Inverse Fourier Transform will be able to restore \( s(k) \).

\[
s(k) = \sum_{u=0}^{N-1} a(u) e^{\frac{i2\pi k u}{N}}, \quad k = 0, 1, 2, \ldots, N - 1 \quad (V-6)
\]

The restored values will be the same as the original points. However, we do not need to use all \( N \) pixel values to reconstruct the original shape. Normally the contribution of higher frequencies to the shape structure is very small; we are able to remove higher frequency components.

So the equation will be as follows:

\[
\hat{s}(k) = \sum_{u=0}^{M-1} a(u) e^{\frac{i2\pi k u}{N}}, \quad k = 0, 1, 2, \ldots, N - 1 \text{ and } k > M - 1 \quad (V-7)
\]

A bigger value of \( M \) results in a shape getting closer to its original (see Figure V-9). In practice we are able to reconstruct the shape reasonably well by using few descriptors.

As previously explained, Fourier descriptors generally smooth out fine details of a shape and hence can be used for template matching applications. One of the advantages of this feature is its independence from location of the edge within the plane. High frequency in image processing refers to rapid changes in the pixel intensities moving along the image. Similarly a shape contour with high frequency content implies large changes in the \( x \) or \( y \) coordinates proceeding around the contour.

Suppose that instead of using all components of \( a(u) \) just the first \( M < k \) points are used, fine details of shape contours will be low pass filtered and just an approximate shape will be reconstructed; by adding additional terms the shape contour will be refined.

![Original](image)

![Figure V-9 Fourier descriptor with different number of M (frequency components)](image)

*Shapes with larger \( M \) are more similar to the original*
V.3.2 Video Compression Implementation

MPEG-2 (the most common block-based standard for lossy compression) is used in this experiment. MPEG-2 is compatible with a wide range of video resolutions from low resolution surveillance cameras (352x288 pixels) to high definition broadcast television (1920x1080 pixels) and up to 30 frames per second. MPEG encoding reduces the redundant video information both within a frame and between frames [200]. In other words, it compresses individual frames by spatial (intra-frame) compression techniques and groups of frames together by temporal (inter-frame) compression techniques. As explained in II.4, Intra-frame coding means that each frame is encoded using JPEG image-coding syntax [201]. The Flowchart of Intra-frame coding is depicted in Figure V-10.

In this scheme, an original image is divided into DCT microblocks consisting of 8x8 pixels. These blocks are transformed by a discrete cosine transform (DCT). Heavy truncation of high frequency DCT coefficients leads to a constant value for many elements of compressed micro-blocks, visible as blockiness.

These artefacts are more visible in smooth areas due to the introduction of sharp discontinuities from one microblock to the next. On the other hand, such micro-block to micro-block discontinuities are less readily visible in areas of highly detailed features and yet have a major distortion impact, which can lead to failure to recognize small objects, such as alphanumeric characters, in forensic analysis of lossy compressed video.

In this work, one intra-coded luminance-only frame is followed by eight predicted frames based on the previous estimated frames. Matlab code (inspired from open-source ffmpeg2 [202, 203]) is implemented in this experiment.

Spatial frequency coefficients using an 8x8 Discrete Cosine Transformation were quantized by the default quantization table in the MPEG-2 standard [204] in a manner identical to that used in JPEG.
image compression. However, the usual implementation of the quantization table in MPEG-2 is to use the default table defined in the standard, and specify a linear scaling factor $q$ to achieve changes in quality. This is described by

$$Y_q(k,l) = \left\lfloor \frac{Y(k,l)}{qQ(k,l)} + 0.5 \right\rfloor, \quad 1 \leq k, l \leq 8.$$  \hspace{1cm} (V-8)

Where $Y(k,l)$ is the un-quantized 2D DCT coefficient matrix, $Y_q(k,l)$ is the corresponding uniformly quantized coefficient, and $Q(k,l)$ is the quantization step size at $(k,l)$.

The first frame in every nine-frame Group of Pictures in our implementation is an intra-coded reference frame using the abovementioned scheme to create a lossy-compressed reference frame. The eight subsequent frames are encoded as a combination of exhaustive search motion estimation vectors and a residual intra-coded error field. The last step was to encode the video frames in an MPEG-2 compliant bit stream. If $q$ is the scaling factor for the quantization table it can be inferred that the greater $q$ the coarser quantization table is in quantization process. In our experiment $q$ is varied from 2 to 23 (with step size of 3) to see its effects on character structure. To compare the impact of compression on the characteristic shape of characters, the cross correlation between compressed and uncompressed character Fourier descriptors were computed. As previously mentioned, the fifteen lowest-order coefficients are retained in order to compute Fourier descriptors in this experiment. Fifteen coefficients were able to capture 80% of the energy in our test and therefore it could effectively show the main structure of each character in the Fourier descriptors.

In the next stage the proposed bio-inspired video enhancement is implemented on uncompressed video and the subsequent video underwent the same compression processing stages. Correlation between Fourier descriptors in enhanced uncompressed video and the output video (that is the compressed version of enhanced video) were also computed to compare to the previous correlations.

### V.4 Statistical Analysis

To examine the results we draw attention to characters which could cause more confusion in highly compressed videos. Therefore, characters having the highest cross-correlation (i.e. C and G; H and N; O and Q) were considered. The boxplot [205], shown in Figure V-11, provides a summary of the results. The band approximately in the middle is the median cross-correlation. In the original (non-enhanced) video sequence, the distribution of correlation coefficients is generally wide and relatively low, as shown in the red/narrow box. As $q$ increases the video quality decreases and so does the mean correlation coefficient. The inclusion of the Large Monopolar Cell pre-processing stage at high quality levels (small $q$) results in a significant increase in mean correlation and decrease in spread of the 50% confidence interval. However, once the video quality is degraded beyond a certain point ($q=14$ in this example) the pre-processing advantage is lost and there is no significant improvement in
the mean and spread of the correlation coefficient. One can conclude that the proposed enhancement system normally does better under moderate to high quality compression. We infer from these results that for compressed video of moderate to high quality, biologically-inspired pre-compression processing to enhance object edges will improve object recognition by increasing correlation, over a narrower confidence interval, to the object template. In turn, this would improve discrimination between ambiguous characters such as C versus G, H versus N and O versus Q, as shown in Figure V-12. To show the statistical significance of the enhancement method, the proportion of enhanced video frames having higher correlation were also calculated (see Figure V-13). Obtained results also confirm that the LMC algorithm results in better performance in lower quantization scales. However, under high levels of compression (q=23) there is some degradation in correlation.

The primary objective of this work was to explore whether spatio-temporal pre-processing of frame-by-frame video, based on a biological model, provides any real advantage for post-compression analysis. The specific application was the recognition of alphanumeric characters, as represented by cross-correlation between objects extracted from a synthesized video sequence and candidate templates. Future work may extend this exploration to include real video conditions.

The biologically-inspired pre-compression processing stage proposed in this work has been demonstrated to improve alphanumeric character structure correlation in compressed video at moderate to high quality. Employed parameters in the LMC implementation have shown convincing results for this experiment. However, optimization techniques can also be used in order to choose better parameters in bio-inspired enhancement methods for other applications. Using such enhancement in surveillance cameras could improve object recognition in a statistically significant way. A variation on the approach described here is to include other biologically-inspired vision stages to detect region of interest (ROI) and specify an appropriate quantization table and thereby ensure that storage or transmission bandwidth is reduced further, subject to quantifiable image quality requirements. Some of these variations have been considered and tested in later stages of this research.
Figure V-11 Characters Autocorrelation comparison- This boxplot shows how the proposed biological pre-enhancement method impacts on character structures obtained from compressed video.

In this box plot middle line is the mean of measured correlation, and the edges are the 95% confidence interval. Circles and stars identify the outliers in this distribution.

Higher autocorrelation suggest that characters structures have been less impacted by compression algorithm.

This enhancement method however will not be helpful in higher quantization scales as the detrimental effects of compression outweigh the impact of LMC pre-compression.
Figure V-12 Separation ability comparison in order to see the biological enhancement impact
The comparison is made between compressed (quantization scale: 15) and uncompressed video when LMC is (or is not) implemented
This boxplot demonstrates separation ability improvement: the increase in autocorrelation with corresponding decrease in cross-correlation can be inferred directly as improved separation ability
V.4.1 Comparison with Conventional Methods

As mentioned before, the primary objective of this research is not to propose a new number plate recognition system. However, it should be acknowledged that there are some effective enhancement approaches proposed in ALPR systems for the purpose of improving recognition system performance. Therefore, the performance of our biologically inspired method is compared with one of the most effective conventional method used in similar applications (i.e. Canny edge detector). In other words, we assume what would have happened if we were using a Canny edge detector as the pre-compression enhancement stage instead of our biological model and if it could lead to even better results. Template matching proposed in [206] is used for character recognition stage in this experiment. Template matching is reliable and straightforward, based on templates which are normalized and stored in the database. The extracted character is matched with all the characters in the database using a Hamming distance approach. The mismatches for each of the extracted characters are measured by comparing with the original characters in the database. The character with the lowest
Hamming distance is regarded as the recognized character. This system struggles to achieve very high recognition rate as the video is noisy.

Table V-2 gives results for the percentage recognition of character segmentation, and character recognition. A failure results if a character is not properly segmented and it is incorrectly interpreted. This normally happens due to poor video quality. It is shown that the system is generally able to correctly recognise around 155 out of 200 test images. The results in Table V-2 show that both methods can improve the system segmentation and recognition accuracy. The biological method however can significantly improve the accuracy of character recognition as it will help the system to retain important structural detail of character even after compression. The biological vision not only enhances the character edges, it also creates a distinctive border between character edges and background. This method is not just spatial filtering process but an exploitation of temporal processing by which characters details are efficiently enhanced. In other words, it would not amplify noises while it is enhancing the character edges; therefore, edges can still be recovered even after compression.

Table V-2 Recognition rate for character recognition.
Character recognition performance, using conventional method, has increased approximately 5.78 % and 18.24% when Biological method is used. These results clearly present the biological method capability in improving character recognition systems accuracy.

<table>
<thead>
<tr>
<th>Total image (240 frames)</th>
<th>Accuracy of Character Recognition (%)</th>
<th>Performance Improvement Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Quantisation Scale =11</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>G</td>
</tr>
<tr>
<td>No Enhancement</td>
<td>75</td>
<td>73</td>
</tr>
<tr>
<td>Conventional Method</td>
<td>78</td>
<td>76</td>
</tr>
<tr>
<td>Using Canny Edge Detector</td>
<td>85</td>
<td>86</td>
</tr>
<tr>
<td>Biological Method</td>
<td>90</td>
<td>88</td>
</tr>
</tbody>
</table>
VI.1 PROBLEM STATEMENT

VI.1.1 Challenges for surveillance systems

Surveillance cameras are widely used in many areas such as traffic control, airport access and monitoring of remote places, warehouses and secured areas. In many of these applications cameras constantly capture video over long periods of time. Video footage captured by surveillance cameras needs to be compressed to meet both storage and transmission constraints. The only global option is to exploit very lossy compression, compromising potentially useful evidence. If useful data could be identified in real time it would greatly assist compression. In fact, most video compression systems are intended for general purpose videos in the context of entertainment rather than capture of forensically valuable detail, hence they are not the best choice for compression of surveillance video. Most surveillance cameras are either stationary or quasi-stationary (such as Pan-Tilt-Zoom (PTZ) cameras) [207]. These cameras usually monitor the same area for months or even years. The background of such monitored scenes is mostly static or has few changes over a long period of time. Hence, Region of Interest (ROI) coding can be exploited, provided that an effective method can be found for identifying regions of specific surveillance analysis interest.

In this section, we propose a novel and efficient biologically-inspired ROI detection method, which is used to identify regions most likely to be of value in surveillance footage, notably moving objects. We are only concerned in this case with determining whether video contains moving objects, and if so, providing detailed and reliable compression where such moving objects occur. In the absence of such a processing stage, inter-frame based video compression treats all parts of the video frame uniformly, resulting in a relatively high quality static background and blurred detail of moving objects. Objects such as faces and license plates are thus difficult to identify in conventional compressed video footage.

The focus of this work is on scenarios in which the monitored area does not have ideal lighting. Despite all research effort in the topic of motion analysis and region of interest coding the literature has not, to our knowledge, addressed the question of whether motion detection systems can still perform well in low light condition surveillance applications and how they can be exploited in video encoding systems in order to reduce the distortions caused by compression algorithms. Much surveillance footage is captured under non-ideal lighting, including both deep shadow and low light which can in turn introduce sensor noise artefacts. It is known that biological video systems do perform well under these circumstances. The proposed biologically inspired spatiotemporal
processing algorithm aims to retain more semantic information for surveillance purposes with noise reduction. Since it makes no assumptions about the encoder, this system is low in complexity and requires no modification to a subsequent video compression system. In this chapter, the literature on Region of Interest (ROI) coding and how it can be implemented in conventional video encoders is reviewed. Experiments and results are then discussed, with the conclusion and proposed future work presented at the end of this chapter.

VI.1.2 Region of Interest Coding

Video compression is associated with data reduction while having as little effect on the quality as possible. This concept has been advanced by employing Region of Interest (ROI) coding. Higher quality within the ROI normally gives the viewer an impression of high quality video without changing the total amount of data. Video sequences with high spatial frequency details and motion contents will be most impacted in low bit rate (i.e. highly compressed) representation. Without suitable region of interest coding important information will be lost. This can become even more important than experiencing poor perceived quality in the resultant encoded surveillance video.

Since the network link capacity is limited in many applications such video coding is able to increase the quality in regions of interest at the expense of reducing quality in the background (when it is compared to ordinary video encoding). Region of interest is determined by application. For example, the face is of interest in human tracking or video-conferencing, and license plates are of interest in monitored parking areas.

The MPEG-2 (and its related predecessors such as MPEG-1) compression standard is aimed at reducing the data rate of digitalized video significantly by exploiting perceptual methods. Specifically, it inserts regular intra-coded frames using down-sampled chrominance and discarding of high spatial frequency detail, along with inter-coded frames based on block matching against reference frames and subsequent residual error. This approach treats the entire video frame uniformly. The concept of region-of-interest (ROI) coding however is introduced to treat areas in a frame unequally and therefore facilitate different strategies for encoding each classified area. Human visual system (HVS) features were considered in [208]. That work showed that the human eye tends to be drawn more on certain areas rather than the whole frame and visual attention is due to some factors including colour, contrast and texture. Thus, macro-blocks receiving more attention are encoded dissimilarly from other macro-blocks. ROI coding methods are also defined in JPEG2000 [209]. In [210], bit and power allocation were also discussed and studied in more detail. Encoding parameters such as quantization, number of reference frames and motion vectors were allocated based on a ROI detection system output. Efficient bit rate control and better subjective visual quality are the main benefits of the abovementioned ROI encoding systems.

Different approaches for ROI bit allocation include controlling encoder parameters or including a pre-processing stage before the encoder. Video compression standards allow the user to alter bit allocation
as long as the syntax of the bit stream is unchanged. Using different step sizes in different regions has
been proposed in [211] and [22]. Smaller quantisation steps for ROI and larger steps elsewhere result
in an improved quality of the ROI, however abrupt changes between ROI and non-ROI quantisation
steps can introduce some disruptively noticeable artefacts in the video. Further, every time the video
storage system is modified, the parameter controls need to be reconsidered. Using a pre-processing
stage instead can avoid these issues. Although the ability to specify ROI quality directly is lost, the
great advantage is compatibility with legacy systems, so that any off-the-shelf video encoder and
particularly an existing encoder in an existing system with predefined operational parameters can be
used. The main approach is to pre-intensify high spatial frequency details in the ROI and thus force
the encoder to retain these details in the encoded bit-stream.

Many surveillance systems use available network bandwidth to transfer captured video to the user.
Therefore, encoders with data rate control are widely used in such systems. Such encoders
dynamically adjust parameters to achieve a target bitrate. A limited budget of bits is the key factor to
determining quantization scales in each frame (or group of frames). In other words, the encoder’s
freedom to allocate more bits is limited by network constraints.

The retention of higher spatial frequency regions in DCT-based video encoding results in more coded
bits and compressing such regions can result in more quantization and prediction errors. Lower
frequency regions on the other hand give more zero frequency coefficients and therefore less error.
Low pass filtering (blurring) of regions in which we have little or no interest can give the encoder
more freedom to allocate the remaining bits to the ROI. In other words, a smaller Quantization Scale
(QS) will be used for frames in which background high frequency details are suppressed and
eventually the number of bits allocated to the ROI will be increased at the expense of a decrease in the
quality of the background. Low pass filtering of non-ROI has been previously addressed in [22] and
[212]. This work adopts the proposed approach in [22] by having graduated degradation in quality
from ROI to the background in order to reduce border effects.

The joint system is comprised of a pre-processing stage and encoder. Having removed higher
frequencies in the background, the encoder automatically allocates more bits, and thus captures more
detail, in motion-rich ROI areas.

VI.1.3 Conventional enhancement techniques

Accurate moving object detection has been a difficult and challenging research problem and has
attracted significant attention in recent years. Background subtraction and temporal differencing are
among the most popular methods proposed in the literature [213, 214]. In background subtraction, a
scene background model is estimated through a learning process and moving objects are detected by
taking the absolute difference between each new incoming frames and a learned background model.
In temporal differencing subsequent frames are subtracted and regions of motion are identified and preserved. This method offers low computational complexity in comparison with background subtraction. However, this method is sensitive to slow changes in the background and further processing is required to remove false alarms. Background subtraction performance can also be adversely affected by illumination and weather changes. There are other proposed methods such as the optical flow method [215] that are computationally complex and cannot be applied to high quality videos in real-time without high-end hardware. The pattern of object motions caused by relative motion between an observer and the scene is used in optical flow methods.

There are also hybrid methods in which previously mentioned methods (i.e. background subtraction technique and frame differencing approach) are combined [5] in order to improve the method performance. To see where the insect motion detection model is positioned, we have compared this model to a relatively effective background subtraction. Piccardi [216] has conducted a review of the most relevant background subtraction method and complexity in terms of speed, memory requirements and also accuracy of different systems are compared.

Amongst the methods reviewed, simple methods offered acceptable accuracy while achieving a high frame rate and limited memory requirements. The review demonstrated that having complex systems would not necessarily lead to better results.

An extensive performance analysis is not possible in the scope of this research, as it would require agreement on an experimental benchmark. Here we limit the discussion to compare the biological motion detection system to one of the simple yet effective background subtraction algorithms. Therefore, the Running Gaussian average is selected as the method that is compared with our proposed method.

In this model, each \((i, j)\) pixel is independently modelled and this model is based on fitting a Gaussian probability density function (pdf) on the last \(n\) samples (i.e. pixels) in time. The running average of the new frame in this approach is computed as:

\[
\mu_t = \beta I_t + (1 - \beta) \mu_{t-1}
\]

\((VI-1)\)

\(I_t\) is the current pixel’s value and \(\mu_{t-1}\) is the previous frame average; \(\beta\) is normally tweaked based on the application - choosing between stability or fast update. The standard deviation \((\sigma)\) of the Gaussian pdf is also updated on arrival of each new pixel. This method requires a relatively low memory requirement as each pixel has only two parameters \((\mu, \sigma)\).
The classification process in this method is described here:

At each frame, the current pixel \( I_t \) can be classified as a foreground pixel if:

\[
| I_t - \mu_t | > k \sigma_t \quad \text{(VI-2)}
\]

And clearly if it does not meet the inequality requirement \( I_t \) will be classified as a background. In this method, the lower update rate can reduce the computational load of the system. However, in that case the system might not be able to respond quickly to the actual background dynamics.

Morphological filtering is one of the main steps in background subtraction methods. The morphological growing technique is implemented to develop a single region and connect all the foreground pixels. A single connected region corresponds to all areas of the moving object.

The main challenge here is to either fill out the region between the set of near connected blobs or consider them as individual objects. Different approaches might be used in the closing operation depending on the assumption of object sizes in the scene.

### VI.2 Elementary Motion Detection (EMD) Model Description

Elementary Motion Detection, shown in Figure VI-1, receives the signal from a visual sampling unit (e.g. one ommatidium), which is delayed and compared with the non-delayed signal from a neighboring unit - or in other words a luminance change at one point on the retina is correlated with a luminance change at a neighboring point on the retina. This motion perception is referred as “first-order” motion perception and is the modified version of the Reichardt motion estimator model. This delay introduces asymmetry to the system and the direction of the motion can be identified by an asymmetrical feature. In simple terms, the results of two correlators are subtracted: a positive output can be indicative of a rightward motion and a negative output indicates a leftward motion, while a zero output shows no left-right motion [217, 218]. Reichardt motion estimator models can function as a reliable velocity estimator. In fact the output of a basic Reichardt correlator provides an accurate indication of image velocity in laboratory tests. Being said, Later studies showed that spatiotemporal pre-filtering can significantly improve such motion velocity estimators [219].

The EMD array is fairly robust to temporal and spatial noise [137]. A two dimensional EMD (2-D EMD) model is used to process the video input which was already spatiotemporally pre-processed by the Photoreceptor and LMC stages.

The delay can be adjusted to select a preferred motion velocity. Multiple parallel EMD outputs are correlated to detect block motion or global motion in the visual field.
Since the 2-D EMD array can generate velocity magnitude and direction, whereas detecting motion is the only task of EMD in our model, only the absolute value of 2-D EMD output is required in this work (see Figure VI-1). We later normalised this absolute value in order to simplify the thresholding process. One of the advantages of this model is that it can enhance the target movement, whereas it suppresses the background movement. The rough position of the moving target can be easily located by finding the peak.

![Figure VI-1 Structure of a Elementary Motion Detector](image)

**VI.3 EMD IMPLEMENTATION**

**VI.3.1 Experiment Description**

To test the proposed system, a high speed radio control (RC) car was used. Since the capture of such a moving object with regular frame rate cameras (25 or 30 f/s) could introduce some motion blur, a USB-3 high frame rate and high resolution camera (FL3-U3-32S2C/M) were used. This is a grayscale camera and only luminance is captured. Another advantage of using such cameras is that we can have access to raw data from CMOS sensors and no other image processing algorithms such as noise removal and contrast enhancement have been applied. The purpose of this experiment was to generate a real condition surveillance video under complex lighting conditions.

In order to simulate poor lighting conditions, directional light beams were generated by a lamp with bendable LED arms in a dark room (see Figure VI-2). No other light source was in the laboratory and LED arms were pointed in different directions. In general, our image acquisition system consisted of a CMOS camera and high resolution machine vision lens (Fujinon HF12.5SA-1) and computer with
image grabbing codes. The abovementioned robust lens is particularly suited for applications where highly detailed close-up imaging is a priority. Moreover, its wide-aperture (F1.4) design makes this lens a very suitable candidate to capture more details under low light situations. The original raw captured video had 1280x720 resolution with 60 f/s of which 10 seconds (600 frames) were selected for study. However; we later changed the display video frame rate to 25f/s in order to be compatible with computer and TV standards. In our video the car started to move forward and backward randomly at maximum speed (approx. 4 m/sec). Due to low light conditions, some noise was introduced to the resulting video. However, alphanumeric characters were still recognizable. Since this experiment was fundamentally designed for character recognition studies in compressed videos, similar characters that can be confused in highly compressed videos (e.g. C and G and O and Q) were picked and placed on the number plate attached to the RC car (see Chapter V).

Insects make the best use of contrast and motion information in visual scenes. To exploit insect vision system features, an accurate model needs to be implemented in moving object recognition systems. Such biologically-inspired vision enhancement has advantages over existing still image enhancement, including high noise immunity. Biological algorithms also work very efficiently in low contrast or luminance situations which are problematic in many automated recognition systems [23].

Robustness to temporal and spatial noise can make this pre-processing algorithm useful in many real monitoring applications.

The most distinguishing feature of EMD (combined with LMC) is its ability to detect moving object structural details rather than albedo features. This feature helps us to detect areas in which more high
frequency data is contained. EMD and LMC parameters were obtained from the Brinkworth and O’Carroll model [10]. A simplified algorithm of the proposed motion detection system is shown in Figure VI-3. Note that three neighbouring pixels are employed in this implementation and the normalised absolute value is used for motion detection purposes.

After thresholding and normalizing the Motion detection model, high pass filtering (i.e. Canny edge detector) and closing operations were carried out to enclose the Region of Interest. Low-pass filters were used in the region of non-interest to remove high frequency data. In order to keep simplicity and avoid disjoint artefacts caused by low pass filtering around the moving object, two low pass filters were implemented in this work.

A rotational symmetric Gaussian low pass filter of size 16 with standard deviation of 2 was empirically determined as the main low pass filter and smaller Gaussian filter of size 8 was manually chosen for neighbouring areas of region of interest. Two corresponding frames in pre-processed and original videos are shown in Figure VI-4.

```
D=Input – k1 ;
HPFFrame= Input - Spatial Low Pass Filter (Input)
LMC1LPFPrevState = (D- LMC1LPFPrevState) / t_constant_1 + LMC1LPFPrevState
LMC1 = D – ( K2 x LMC1LPFPrevState)
LMC2LPFPrevState = ( HPFFrame – LMC2LPFPrevState) / t_constant_2 + LMC2LPFPrevState
LMC_Output = tanh ( k3 x (LMC1 – LMC2LPFPrevState))
```

```
Neighbour (1) = LMC_Output (row-1 , col+1)
Neighbour (2) = LMC_Output (row , col+1)
Neighbour (3) = LMC_Output (row+1 , col+1)
```

```
EMDPrevState = (LMC_Output - EMDPrevState) / t_constant_3 + EMDPrevState1
EMDNeighbourPrevState = Neighbour (1/2/3)-EMDNeighbourPrevState1)/t_constant_3 +
EMDNeighbourPrevState
NeighbourUp(1/2/3)=tanh ( EMDPrevState x Neighbour (1/2/3))
NeighbourDown(1/2/3)= tanh ( EMDNeighbourPrevState x LMC_Output)
```

```
Result = abs (.7 x (((NeighbourUp(1) – NeighbourDown(1))+(NeighbourUp(2) - NeighbourDown(2))+(NeighbourUp(3)-
NeighbourDown(3))))
```

**Figure VI-3 Implemented EMDs (combined with LMCs) Algorithm**

**Three EMD algorithm are carried out for horizontal, vertical and diagonal neighbors and the resultant will be computed in the last stage – In our experiment k1=.4, k2=.72, k3=2, t_constant_1=22, t_constant_2=12, t_constant_3=4 (note: The constants are obtained from Brinkworth model [10 ] and subsequently modified to be compatible with the application)**
Figure VI-4 Top frame is the original captured frame. In this example compressed frame (without pre-processing) cannot be easily read whereas pre-processed and compressed frame still provide good amount of details for character recognition purpose (note: frame contrast is further enhanced for demonstration purposes in these images).
This picture clearly demonstrates how high frequency details in background are suppressed by low pass filtering. To test our pre-processing algorithm, the MPEG-2 library (mpeg2video) in ffmpeg was selected to encode our videos [202, 203]. This popular framework supports most codecs designed by the MPEG standards committee. Its interface offers some freedom to change the encoded video quality and data rate. Different average bit rate with same Group of Picture (GOP) structure (i.e. IBPBPBPBI) videos are tested in this experiment. The next stage of the experiment examines the impacts of pre-processing stage and compares the encoded video with and without it. The most important objective of the experiment is to ensure that the implemented encoder does not have a significant impact on information existing in the original video. Otherwise the fidelity of the video can be legitimately questioned. PSNR is the most common quality measure in compressed videos. Note however that since blurring the background will be regarded as noise in the signal, comparison is performed between the ROI of pre-processed video and encoded video (see Figure VI-5).

Figure VI-5 PSNR is compared between the reconstructed video after compression and the video after spatio-temporal processing.
There have been many cases where investigators have obtained positive identification using surveillance footage and many court decisions that have allowed video and image into evidence [220]. Hence, reliability of such videos can play a crucial role in many cases. As previously mentioned, Peak Signal to Noise Ratio (PSNR) has been widely used as an objective video quality metric to evaluate the proposed system performance. PSNR of segmented object (i.e. number plate) with and without pre-processing is initially compared as shown in Figure VI-6. The video with pre-processing has approximately 2 dB higher PSNR showing a significant improvement in terms of compressed video fidelity. Although PSNR is relatively a weak metric to compare original and compressed structural details, this graph clearly demonstrates that less details are altered in ROI. Further investigation is still required to measure the effectiveness of the proposed pre-processing in a particular application (i.e. character recognition in this research).

![Figure VI-6 Segmented Object PSNR comparison of Encoded Video (MPEG-2) with and without pre-processing](image)

*Note: This is not PSNR between whole frames*

- PSNR 95 %confidence interval (without preprocessing) = [25.694, 25.743]
- PSNR 95 %confidence interval (with preprocessing) = [27.726, 27.755]

In these cases more than 2 dB improvements has been made in the PSNR. These results demonstrate the effectiveness of biological enhancement in retaining useful information.
In the next step, the autocorrelation of individual label characters are also compared. Autocorrelation is used to combine together all parts of the image in order to find repetitive structures. Therefore, altered structural details due to motion compensation and quantization artefacts can be easily identified by comparing autocorrelation values.

Exhaustive comparison is individually made to examine the above-mentioned altered high frequency details. Encoded videos with different data rates are also tested to examine the PSNR and autocorrelation behaviour for different network bandwidths (see Figure VI-7 and Figure VI-8). Since a very low data rate can result in distorted number plate (difficult to recognize), 256 kbps is set as the lowest data rate. It is observable that both PSNR and autocorrelation has been improved in the produced video. Curiously, as the date rate increases the autocorrelation gap narrows, while the PSNR gap widens.

![Autocorrelation and PSNR comparison](image.png)

*Figure VI-7: PSNR and Autocorrelation comparison for different data rate
these graphs clearly shows the difference between autocorrelation and PSNR as video quality metrics. For instance, character structures will not be changed significantly in higher data rate. However, this does not necessarily mean that there is no room for significant signal to noise ratio improvement (e.g. approx. 4dB PSNR improvement in 600kbps versus only .015 auto-correlation improvement)*
This demonstrates that PSNR by itself is not a very useful metric to measure the effectiveness of video enhancement method. We also can conclude that the proposed algorithm becomes even more useful in medium range data rates when it seems no more improvement is required.

VI.3.1.1 COMPARISON WITH CONVENTIONAL TECHNIQUES

The main goal of this research is not to propose a new number plate recognition system. However, it should be acknowledged that there are some effective enhancement approaches proposed in ALPR systems in order to improve the recognition system performance. This experiment is fundamentally conducted to investigate how video compression methods can be modified by which even small details on moving objects are recognisable. Therefore, in this section the performance of our biologically inspired method is compared with one of the most effective conventional motion detection method used in similar applications. Here we limit the discussion to compare our proposed biologically inspired motion detection system to the Running Gaussian average background subtraction method. 240 frames are randomly selected from our experimental video. These two
methods are used in order to segment moving objects from background in the ROI based video compression (discussed in VI.1.2).

In simple terms, we have used the same background smoothing and lossy video compression techniques and the template matching [206] is also used (same as V.4.1) for this character recognition stage (see Figure VI-9).

The well-known criteria are used to evaluate the performance of the proposed algorithm against the conventional method. Precision ($P$) and recall ($R$) are used here in this experiment.

\[
\text{precision}(P) = \frac{\text{Number of correctly detected segments}}{\text{Total number of totally detected segments}} \quad \text{(VI-3)}
\]

\[
\text{recall}(R) = \frac{\text{Number of correctly detected segments}}{\text{Total number of ground truth segments}} \quad \text{(VI-4)}
\]

Figure VI-9 flowchart of comparison method (Proposed vs. Conventional)

This flowchart shows how the conventional and proposed method are compared in moving object segmentation stage
In order to provide the ground truth data set we need to find out what type of ground truth is needed and how to generate such ground truths. In this case, it is simply expected to segment an image into two regions (i.e. foreground and background). Therefore, ground truth data set are manually generated by segmenting moving objects in each frame. In fact noisy and dark environment makes it impractical to use other existing methods. The performance of the proposed method and conventional method for the 240 frames is compared and shown in Table VI-1. This table shows the efficiency of each system in segmentation detection.

A comparison between the proposed algorithm and a background subtraction algorithm (running Gaussian average method) was performed, which demonstrates that the using proposed algorithm can offer approximately 30% higher precision and 29% higher recall. In this experiment the correctly detected segment is the one occupying at least 85% of ground truth segment area that has been manually detected by human user.

<table>
<thead>
<tr>
<th>Total frames</th>
<th>number of correctly detected segments</th>
<th>number of totally detected segments</th>
<th>Precision $(P)$</th>
<th>Recall $(R)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conventional method</td>
<td>240</td>
<td>142</td>
<td>206</td>
<td>69%</td>
</tr>
<tr>
<td>Biological method</td>
<td>240</td>
<td>182</td>
<td>209</td>
<td>87%</td>
</tr>
</tbody>
</table>

Table VI-1 Overall performance comparison (Conventional vs. biological)
Approx. 30% precision and recall improvement when biological method is used (Gaussian average method is employed for conventional method)
In Table VI-2 the recognition rates between conventional and biological model are compared. Although higher compression could have a severe impact on the recognition performance, moderate data rate can still demonstrate the overall performance of the character recognition system under different enhancement methods. It should be acknowledged that the effects of other varying parameters (e.g. light intensity, character size etc.) can also have some impacts on the recognition system performance. However, investigating their influence in this experiment could introduce some complexities to this study which can distract us from the main purpose of this chapter (i.e. implementing insect vision system model in ROI coding). In further studies, not only the individual parameters impacts should be studied but also their combined influences need to be investigated. However, this exhaustive testing is outside of this research scope.

<table>
<thead>
<tr>
<th>Total image (240 frames)</th>
<th>Accuracy of Character Recognition (%) Avg. Data Rate = 300 Kbps</th>
<th>Accuracy Improvement (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No enhancement</td>
<td>C 58, G 56, H 57, N 63, O 52, Q 49</td>
<td>12</td>
</tr>
<tr>
<td>Conventional Method</td>
<td>64, 63, 65, 68, 60, 57</td>
<td></td>
</tr>
<tr>
<td>Biological Method</td>
<td>85, 81, 86, 87, 85, 82</td>
<td>49</td>
</tr>
</tbody>
</table>

*Table VI-2 Character Recognition Results*

In our experiment biological method provides 37% more recognition accuracy compared to conventional method.
Chapter VII  INTEGRATION – EARLY STUDY ON FACIAL RECOGNITION

VII.1 INTRODUCTION

The goal of this chapter is to introduce existing challenges in human and machine recognition of faces and how the proposed insect vision model can be implemented to improve accuracy of face recognition tasks.

Automated recognition of faces has different applications, including mug shots, card verifications, or monitoring individuals in secured areas by surveillance cameras.

Most of these applications have to operate under a set of different constraints in terms of processing complexity and quality requirements, thus creating a wide range of technical challenges. Researchers in the computer vision and image processing area have examined a number of issues related to face recognition by human and machines over the last 30 years. Proposed methods have been tested on different sets of images with different features. However, little research exists in investigating impacts of the different processing stages of the video acquisition system on face recognition tasks by humans or machines. Most importantly, there is no benchmarking study to determine certain circumstances under which face recognition performance cannot be significantly affected. In fact, the larger the margin of error, the less confidence we would have that the employed system’s reported results are close to the truth. Quality is arguably even more important in law enforcement applications.

In this chapter, various applications of face recognition in law enforcement and commercial sectors are presented, followed by a summary of the wider literature on face recognition. An overview of face segmentation, feature extraction, and recognition techniques is also presented.

One of the most challenging problems in face recognition is detecting and recognising individuals in cluttered scenes with complicated lighting conditions, captured by surveillance cameras.

Several issues are discussed that can potentially have significant impacts in recognising a person in surveillance video footage. This chapter is particularly allocated to face recognition and challenges under complicated circumstances.

Given the numerous theories and techniques that are applicable to recognition tasks and various types of surveillance acquisition and compression systems it is desirable to evaluate different pre-processing techniques by which recognition rate can be improved.

Automated face recognition systems have become an active research area in recent years. They exploit image processing, computer vision and pattern recognition techniques and neural network concept.
There is a considerable difference between how humans recognise faces in cluttered scenes and how machines perform this job. Obviously humans are more capable to do such tasks and it is a much more daunting task for automated systems. Understanding how the human vision system perceives faces and connecting it to already used face recognition techniques and systems can open a new research area. Human faces normally occupy a small portion of the surveillance camera’s field of view. Since all facial features are represented by a limited numbers of pixels in poor resolution cameras, this can even worsen the scenario and raise issues in face recognition tasks. It can be a real challenge to a computer or even a human operator to accurately identify an individual from the available data-base (see Figure VII-1).

The chance of computer and human error in identification tasks is normally higher in low quality videos. One solution to the problem is to enhance the captured video quality by applying different image processing techniques, and hence improve human ability and computer performance in recognising faces. One downside of this solution though is the adverse effects of noise, which can be

*Figure VII-1 this is a typical example of available surveillance video and even magnified view of human face does not include enough details to identify the individual in the scene (this video frame is captured by Wireless IP Surveillance Camera (NEO Coolcam))*
CHAPTER VII INTEGRATION- EARLY STUDY ON FACIAL RECOGNITION

introduced during enhancement techniques.

The second and probably more effective solution is to implement an intelligent system in which faces or regions that are most likely to be used for further analysis receive more attention from the acquisition and compression system. Therefore, we can provide more information to automated face recognition systems (or human operators). This approach is aimed at reserving high spatial frequency details in facial areas that would be lost through the conventional compression process. Such high frequency details can be even temporally and spatially amplified, and hence the face recognition system performance can be improved.

It has been observed that different types of artefacts are usually generated in the acquisition and compression stages. These artefacts can be visually distracting or affect machine recognition algorithm accuracy. There have been a few studies that have shown the impacts of image resolution on recognition accuracy [221, 222]. Super-resolution, for example, would be one of the efficient solutions to improve image quality. However, further studies are required in order to measure the effectiveness of such approaches in machine recognition algorithms. In fact, there is a lack of systematic studies on the automated face recognition systems performance under different surveillance video conditions.

This chapter aims to clarify the gap between face recognition and video encoding systems in complex lighting conditions, while suggesting solutions inspired by biological algorithms with some preliminary experiments.

Automatic face recognition has been the subject of studies for a long period of time [223]. However, high efficiency face recognition techniques were developed in the late 1980s and 1990s. Face recognition systems are normally comprised of different stages, of which face detection is a necessary early step in most of them. Faces in surveillance videos will have different poses and expression. They are also located in different points of camera field of view and this high degree of variability can make face-detecting tasks more complex than other pattern detection issues. Automated face detection is commonly used in surveillance systems for access, control, and monitoring secured areas.

Face detection in video brings new challenges compared to face detection systems based on static images. In fact, real-time video can provide valuable spatiotemporal information, which will not be obtained via static images. Face tracking and recognition tasks can be completed by using this information. Face detection techniques have been under study for a long time and different methods have been proposed in the past decades. One can classify the proposed approaches into feature-based and statistic-based methods. Skin colour, face contours, or other face geometric characteristics are used in feature-based methods [224, 225, 226]. The Adaboost algorithm [160], neural network [227, 228] and support vector machine [229] are included in statistic-based approaches.

Simplicity is the main feature of feature-based approaches. However, their performance can degrade in noisy videos. Occlusion, lighting conditions, and changes in facial expression can also have some
CHAPTER VII INTEGRATION—EARLY STUDY ON FACIAL RECOGNITION

influences on these methods. Statistic-based approaches, however, are more tolerant to noise and distortion.

Some of these systems can handle different scenarios, specifically when there are changes to faces in relation to scale and rotation [230]. Many of these methods are based on face detection in photos. A combination of SVM model and Haar-like feature were used in the Cuong Nguyen Khac et al. model [231]. The downside of statistics-based methods is their complex structure that makes them computationally expensive.

In the vast amount of available face recognition techniques, boosting [232, 233] and PCA-based methods [234] are very popular in real surveillance applications. In the AdaBoost technique, several weak classifiers are linearly combined to develop a strong classifier. By using integral images (also known as a summed-area table), the classifiers are designed based on weighted training error. In other words, larger weight will be assigned to classifier with smaller error and accordingly smaller weight will be assigned to classifiers with larger error. Although object detection using AdaBoost can be slow in the training phase, but it works well in detection step even on resource limited systems.

VII.2 FACE RECOGNITION AND COMPRESSED VIDEO

Although there has been extensive ongoing research to improve the real-time performance of these approaches, they cannot be implemented on high quality cameras in which large data handling is required.

Some researchers have also tried to combine different technologies to make use of their advantages, and hence promote the accuracy of detection. Some have also applied different techniques to develop better feature analysis and hence lowering the false alarm rate. [235, 236].

The AdaBoost [237] algorithm has low false alarm rate and is thus one of the most robust methods proposed in the literature. This method has also shown impressive performance in real time. The system accuracy however, will drop significantly when some distortions (i.e. sensor noise, artefacts due to lighting conditions etc.) are introduced to the captured video.

Other video features can be used to aid improved performance of the face recognition system in noisy conditions. Some approaches for example, have used colour segmentation to remove the non-face background. Such approaches can be ineffective when the background colour is not accessible (e.g. grayscale video) or the colour is similar to that of skin.

A reliable face detection system should be able to detect almost all human faces irrespective of video quality or any complex conditions in real time.
Contrast enhancement is normally the first step in most face detection and recognition methods. Feature extraction and classification are generally the subsequent stages in which different techniques and approaches can be used. The Gabor wavelet filter for example, is widely used in extracting face local features. The typical face recognition diagram is shown in Figure VII-2.

Measuring automatic face recognition performance is a complex problem and the majority of such systems operate within constrained environments. For example, they mostly assume that the subject face is either static or moving with simple transformations (i.e. affine or projective transformation). Global parametric motion (i.e. rotations, translations) is sometimes used to extract the face features [238]. It is common to assume that faces are turned almost toward the front and the direction of the face can have considerable impact on face recognition accuracy. While many of these systems perform well with static scenes, their performance starts to degrade severely when human faces are non-rigid and non-planar, or located in low light areas.

Video and image encoders have long been used to compress raw data captured by camera. There has
been significant effort to identify techniques suited to accomplish this task in different applications, including compressing medical and fingerprint imaging in an optimal manner.

Lossy compression algorithms are employed in order to maximise the compression benefit by significantly reducing the data size. However, lossy compression will introduce some distortions, which can interfere with extracted features that are used in recognition algorithms, hence degrading its results. In other words, the distance between actual and measured face features will be increased in compressed data and this in turn affects acceptance of the recognition system under consideration.

In this chapter the focus is on the impacts of compression on face images and how it affects face recognition performance. In contrast to the majority of studies, the focus will not be on obtained objective and subjective surveillance video quality after compression. A simple yet efficient face recognition system will be implemented to the compressed image data to determine significance of compression effects on recognition accuracy.

There are regions of interest in surveillance videos that are dominated by high frequencies (e.g. number plates). Such regions are more affected by the lossy compression process (since higher frequencies are discarded in frequency domain transformation and quantization). In comparison to these regions, face images have low and medium frequency content as well. Therefore, compression algorithms that are designed for high frequency data (e.g. finger prints compressions) will not be suitable. There are not many compression techniques proposed in the literature that are specifically designed to compress image and videos for face recognition purposes. In the Li et al model for example facial texture images are partially recovered from reference face pictures by least-square minimization [239]. However, this method’s efficiency has not been evaluated under a different range of outdoor and indoor conditions. JPEG and JPEG 2000 compression algorithms have been widely used in most surveillance applications. The effects of these compression algorithms on twelve different face recognition systems were studied by Delac et al. [240]. The JPEG2000 trade-off points between compression ratio and recognition accuracy were determined in work by McGarry et al. [241].

**VII.3 Methodology**

Many different face recognition systems have been proposed and used in industry. This is a preliminary study to test the biologically-inspired pre-compression system; therefore, common methods have been selected for this experiment. We are not going to compare and evaluate different face detection systems’ performance under different conditions. Rather, the aim is to see whether the systems’ accuracy can be improved by using our proposed biologically inspired enhancement method. For this reason, we have chosen a straightforward yet effective, conventional face recognition system.
The automatic face recognition system normally attempts to find the identity of a given face image (or images) according to a database containing training sets. These training sets consist of the features extracted from known face images of different people. The most similar feature vector amongst the training set is selected as the recognized face in this system. In other words, the test image (unknown face) is fed into the system and a feature extraction algorithm will be used to compute its feature vector and compare it to the existing feature vectors in database.

In this experiment, Principal Component Analysis (PCA) is chosen as the feature extraction approach. PCA is one of the most popular techniques that have been used in many applications, including image recognition and compression. The main principle of PCA is to construct a 1-D vector from a large 2-D facial image, which is also known as Eigen-space projection. The implemented PCA algorithm is described later in this chapter.

In this work, a USB-3 high frame rate and high-resolution camera (FL3-U3-32S2C/M) was used to capture a video with poor lighting conditions. The purpose was to generate an indoor surveillance video under complex lighting conditions in which the suspect’s face is completely visible but partially darkened due to side illumination (which is very common in surveillance videos). The original raw captured video had 1600x1200 resolution with 60f/s of which 10 seconds (600 frames) were selected (see Figure VII-3). In our video, the suspect is running towards the camera and then backwards randomly. Since different poses can complicate face recognition tasks (pose-variation is another study case and is out of this work scope), an attempt was made to avoid scenarios in which the facial pose varies. The 3D model is commonly recovered in order to render novel poses. Note that recovering a 3D model from 2D images is not a simple task [242, 243].

The proposed pre-processing stage (i.e. combination of LMC and EMD) is also implemented to segment a moving object to ensure the used face detection system does not need to deal with background noise, and hence its performance can be significantly improved. Many face detection systems make use of human skin colours to detect faces [244]. However, we only have intensity frames in this experiment and we cannot take advantage of the colour characteristics of the skin colour. We are only able to detect a face in an already segmented object (i.e. a person of interest). This stage is a relatively straightforward task as a face is normally located at top of the object. Edge detection (i.e. Canny edge detector) and grey projection and thresholding were also performed in order to efficiently isolate the face. This in fact is a simplified version of employed algorithm in the Baskan et al model [245]. Image pre-processing can play a very important role in the facial feature regions location process, and hence system accuracy can be improved. The main advantage of the photoreceptor in the insect vision system is its illumination adaptability and this can be beneficial, particularly when lighting conditions are not ideal in the scene. In other words, illumination compensation can be carried out on the input video sequence by using the photoreceptor stage by
which deteriorating effects of lighting conditions are alleviated or even completely eliminated (resulting faces after being pro-processed by photoreceptor are shown in Figure VII-3).

In the next stage of this work, MPEG2 was used to encode the captured video. It is assumed that lossy compression algorithms were used to encode the raw video captured by surveillance camera sensors and the resulting compressed video has been used as input to the face recognition system.

We have used general purpose compression algorithms on face image video frames to study the impact on the recognition accuracy of a face recognition system. In the subsequent experiment, the proposed biologically inspired pre-processing stage was exploited to determine if it provides any advantage for the face recognition system.

In this project the Spaceck face database [246] was used, which contains two different images of each of the eleven subjects, including the person appearing in surveillance video (see Figure VII-4). The face pictures used in the training set are centred and scaled with 180x200 resolution.

Figure VII-3 The Video frames captured for this work (Resulting face after being pre-processed by Photoreceptor is shown on top right of each frame)
VII.3.1 Introduction to PCA Algorithm

Principal Component Analysis (PCA) was first proposed by Kirby and Sirvich, who succeeded in showing that PCA is an effective compression scheme by which the mean squared error between the original image and its reconstruction can be well minimized [247, 248]. Turk and Pentland made a significant contribution by establishing a framework for the PCA model [249]. In their work, PCA was exploited to compute a set of subspace basis vectors (also known as “eigenfaces”) for a database of face images. Eigenfaces are actually a set of facial ingredients that have been derived from statistical analysis of many face pictures (our database eigenfaces are shown in Figure VII-5).

In this experiment a face recognition system using the PCA algorithm is implemented. PCA implementation is explained in more detail in this section.

Assume that we have $p$ training images: $\Omega_i, i = 1, 2, \ldots, p$.

The next stage is to form a pixel vector $\phi_i$ (where $\phi_i \in \mathbb{R}^k, (k = M \times N)$) and then compute feature vectors $\omega_i$ (where $\omega_i \in \mathbb{R}^d, (d \ll k)$) from the pixel vector.

A training data matrix $A$ is now generation. Matrix $A$ contains $p$ rows and $\omega_i$ are stored at each row. Thus $A$ has $p$ rows and $k$ columns (i.e. dimensionality of $p \times k$).
The covariance matrix of $A$ ($C_A$) is now computed and eigenvalues and their corresponding eigenvectors of $C_A$ will be obtained. Hence there will be eigenvalues and eigenvector pairs where each eigenvector is $e_i$ of dimensionality $k$. The eigenvalues need to be sorted in decreasing order and the largest $d$ eigenvalues and eigenvector pairs are selected. To compute $\omega_i$ we need to form the transformation matrix $\Psi$ by placing the selected eigenvectors as its column and $\omega_i$ will be obtained simply by: $\omega_i = \Psi^T \phi_i^T$. 

Figure VII-5 Eigenfaces that are derived from training set
In this equation $\Psi^T$ and $\phi_i^T$ are the transposes of $\Psi$ and $\phi_i$ respectively and each column is of length $k$ in which eigenvectors are placed. The dimensionality of this matrix is the same as the input image $(M \times N)$. Therefore, each eigenvector can be converted to an image (also known as an eigenface) by a reversing operation. These images are identifiably similar to human faces.

The main reason for applying this algorithm is to reduce dimensionality (i.e. $d << k$). Thus we just use the largest $d$ eigenvectors to reconstruct the image of person $\Omega_i$. The reconstructed image $\hat{\Omega}_i$ will be a good approximation of the actual image $\Omega_i$. The reconstructed image has been obtained by converting the pixel vector $\hat{\phi}_i = (\Psi \omega_i)^T$ to an image with dimensionality $M \times N$. Note that the more eigenvectors are used, the more similar it becomes to the original picture, which could be fully reconstructed by using all $K$ eigenvectors.

The next stage is to match the test images against images in the database by basis vector projection. In this study there are $N$ possible candidates and the feature vector of each face image is extracted in the training set. In our experiment $\Omega_A$ is the training image of person A with resolution of $M \times N$ ($M=200$ and $N=180$ in our experiment). The image is firstly converted into a pixel vector $\phi_A$ by concatenating each of the $M$ rows into a single vector and the length of the vector $\phi_A$ will be equal to $M \times N$. Vector $\phi_A$ will be transformed to $\omega_A$ with lower dimensionality of $d$ ($d << M \times N$) using PCA. For each training image $\Omega_i$, one feature vector $\omega_i$ will be calculated and stored in a database. In the recognition stage, the test image of a known person $\Omega_j$ (unknown to system) with identity of $\alpha_j$ will be given to face recognition system. The feature vector of this person $\omega_j$ is firstly obtained using PCA and in order to identify $\Omega_j$ similarities between $\omega_j$ and all of the feature vectors $\omega_i$ in the training set will be computed. The similarity between feature vectors is computed using Euclidean distance. The identity of the most similar $\omega_i$ will be the output of the face recognition system. If this identity $\alpha_i$ is the same as $\alpha_j$, it means that we have accurately recognized the person $j$, otherwise it means that person $j$ cannot be recognized correctly. This recognition process is depicted in Figure VII-6.
CHAPTER VII INTEGRATION- EARLY STUDY ON FACIAL RECOGNITION

VII.4 RESULTS

In the first stage, detecting and recognising the suspect face based on the training set was the primary focus. As it is explained in VII.3.1, the test image of a known person (unknown to the system) was given to the face recognition system and if the system identity was the same as the known person’s identity it meant that the suspect had been recognised correctly. If a different subject was identified it meant that the system was unable to recognise the suspect correctly. Note that frames in which faces are correctly segmented (i.e. around 87% of all frames) were only used for this recognition task. Figure VII-7 shows the face recognition accuracy for all compression ratios. The file size of the compressed video was used to calculate the compression ratio of the encoder algorithm (i.e. ratio between raw video and compressed video).

Figure VII-6 Schematic of PCA face Recognition Algorithm
The relationship between the Euclidean distances of the PCA-based recognition system and compression ratio is also studied in the next stage (see Figure VII-8). This figure demonstrates the maximum Euclidean distance between features of test image and correct candidate in training set. These experiments enable us to determine the usefulness of the pre-processing stage in the face recognition and compression context. These figures clearly show that the face recognition system under consideration is quite susceptible to errors when it comes to highly compressed videos. Our proposed bio-inspired pre-processing stage however, could stabilise the face recognition performance in such circumstances.

*Figure VII-7 PCA face Recognition Accuracy vs. Compression Ratio (ratio between the uncompressed size and compressed size)*

Results shows that video compression will benefit more from bio-inspired pre-processing in higher compression rate videos.
**Figure VII-8** Maximum Euclidean Distance (between test image and correct candidate) Vs. compression ratio (ratio between the uncompressed size and compressed size)

*Bio-inspired model can stabilize the Euclidean distance in higher compression rates (note: high Euclidean distance can significantly impair the face recognition system accuracy)*
Chapter VIII  CONCLUSION AND FUTURE WORK

VIII.1 DISCUSSION OF CHALLENGES

As we explored the feasibility of an employing a novel biologically inspired video enhancement method in surveillance video compression, we encountered several issues that had to be addressed. Not only did each component of the biological vision system have their own drawbacks and advantages, but also they were introducing new features when used in combination. We therefore had to add or remove each component based on the nature of application and environment.

VIII.1.1 Surveillance video compression

It has been demonstrated that there has been a gap between some of the apparently relevant topics including video surveillance, biological vision system models and video compression, and a strong link has to be established to interconnect forensic video analysis and the way video compression algorithms are designed. In fact, the video compression algorithms employed in surveillance industry are designed for entertainment purposes and that can make compressed surveillance videos an unreliable source for further forensic analysis. In other words, retaining vital information is the main priority in video surveillance, rather than a pleasant viewing experience.

VIII.1.2 Biological vision system

Video compression designers have exploited human eye features to discard some of the data that either cannot be seen or perceived. This approach however cannot determine the relevance of discarded data. Therefore, a deeper understanding of the human vision system is required to devise more efficient video compression algorithms that are specifically tailored for surveillance video applications. Although we may not be able to employ all human vision features in surveillance due to its complex nature, we may still be able to make use of a similar biological vision system that offers less complexity. The insect vision system has great ability to detect and follow moving objects in complex situations with much lower computational power in comparison with human vision system.

VIII.1.3 Insect vision system and its components

As previously mentioned, many techniques used in current video technology are inspired by the human perception physiology. Human vision is a quite complicated system. Other biological systems however can offer less complexity and they are still very beneficial to real video applications. The insect vision system has an outstanding ability to adapt to different lighting conditions in order to enhance and encode the information captured by the eye. Therefore, insect vision system features were exploited to enhance surveillance videos by which we can retain more information in highly
CHAPTER VIII CONCLUSION AND FUTURE WORK

compressed videos. The obtained results in this research show significant improvement in character
and face recognition accuracy when bio-inspired enhancement methods are employed.

An investigation has been conducted into the problem of video encoding in video surveillance. This
has involved a comparison of conventional encoding systems, which is not specifically tailored for
video surveillance, and a new video encoder in which the proposed bio-inspired pre-processing
system is employed. The tested video encoders are block-based encoders and the main purpose was to
minimise video artefacts caused by such compression methods on regions of interest. Although
different regions of interest might be defined for every video surveillance scenario, the focus of this
investigation was on areas that receive more attention from forensic investigators. Since the human
vision system is more sensitive to low frequency details, block based video encoders discard high
spatial frequency data in order to reduce its size. However, not all of the discarded data is without
useful qualities, particularly in surveillance applications. Alpha numeral characters, structural details,
and facial features are among the most important information that is likely to be lost in such
compression methods.

Complex lighting conditions in surveillance systems has also been studied in this research work as it
can impose deleterious constraints on video quality. A combination of distortions caused by the video
encoding process and low lighting conditions can have considerable deteriorating effects on
identification tasks.

Video quality has a different definition in surveillance video. Therefore, we have had to address the
concept of video quality in a surveillance context, in terms of improved recognition rates by
classification systems rather than a subjective pleasant viewer experience.

In this research work, a spatiotemporal pre-processing method is proposed that is inspired by the
insect vision system. The moving region of interest can be segmented from the background using this
method. Discarding high frequency details in the background gives more freedom to the encoder to
retain more high frequency details in the regions of interest and it has been demonstrated in this work
that this clearly leads to surveillance video quality improvement.

Different objective video quality metrics were computed and compared, leading to the demonstration
of structural detail variation during the encoding process when the proposed bio-inspired pre-
processing stage is implemented.

VIII.2 FUTURE DIRECTION

The primary objective of this work was to explore whether the implementation of a spatiotemporal
biological model as a video processing stage provides any real advantage for motion detection tasks
and region of interest coding in videos captured by stationary surveillance cameras under complex
lighting conditions. The proposed biologically-inspired pre-compression processing in this paper
demonstrates its ability to improve the PSNR in region of interest and also retain more information in such region which was measured by autocorrelation.

Using arbitrary parameters of an Insect Vision Model implementation resulted in convincing results in terms of objective quality requirements. However, optimization techniques can be used to choose better parameters and hence improve the system performance. The efficacy of this method can be evaluated in other relevant surveillance applications by running more authentic experiments.

Low dynamic range video (8 bit) was used in our experiments and a variation on the approach described here is to use high dynamic range video (e.g. 16 bit) and include other biologically-inspired vision stages such as a photoreceptor model to enhance temporal contrast by which motion detection is further improved.

Adaptive high pass filtering features of LMC has considerable potential to assist the human eye to detect small moving objects. Therefore, adding the LMC output to the region of interest can be looked at as the next step of this work. The proposed moving object detection algorithm can also be implemented in object-based (sprite) video coding algorithms addressed in MPEG-4 [250].

Low computational complexity and highly parallel implementation makes this system a suitable candidate for other motion detection applications. The author believes that since the proposed system is designed to detect motion in complex lighting conditions in high noise, it will be also able to deal with other tampering affects appearing in videos used for other applications. However, further experiments needs to be conducted to measure its performance.

Future works could include:

- Improvements to the parts of the method where temporal processing is included to minimize the problem associated with detecting shape boundaries in noisy video.
- Improve the methods for detecting the ROI for face recognition tasks. A variety of videos with different lighting conditions and quality levels can be used in order to optimize system parameters.
- Applying ROI video coding strategies to object video coding.
- Perform additional subjective and objective tests with other bit rates and additional sequences to obtain more reliable results and to examine when jerky movement in the background has a substantial effect on perceived quality.
- Even though some tests are performed on face recognition, it is necessary to conduct extensive experiments. In addition, other applications could be considered where methods of detection have not been thoroughly researched at the present time.
- Incorporate the proposed pre-processing stage into the codec and adapt the rate-distortion optimization based on the results obtained from video quality metrics.
A.1 Filtering in Spatial Domain

Spatial filtering is a principal tool in image processing and it is widely used for a broad spectrum of applications. New pixels will be created by performing predefined operations (based on their neighborhood) on image pixels. Linear spatial filtering for example is given by the below expression:

\[ g(x, y) = \sum_{j=-a}^{a} \sum_{k=-b}^{b} w(j, k) f(x + j, y + k) \quad (A-1) \]

where \( x \) and \( y \) are varied so that each pixel in \( w \) visits every pixel in \( f \).

A.1.1 Canny Edge detector

One of the fundamentals in spatial processing is finding edges in an image in which areas with strong intensity contrast (i.e. large variation of pixel intensity) are extracted. This way some useless data is filtered out while structural details of image are preserved. Sobel, Laplace and Canny edge detection algorithms are popular methods in image processing of which the Canny edge detector is known as the best performing edge detector.

Before implementing the Canny edge detection algorithm, the image needs to be smoothed to suppress existing noise in the image. Gaussian filtering is normally used as a low pass filter due its simplicity. In this step, a suitable mask will be calculated and low pass filtering will be carried out by using standard convolution methods. Convolution masks are usually smaller than the actual image. Large Gaussian masks are not very sensitive to the image noise. In other words, more localization error is introduced if the Gaussian width is increased. They are also computationally expensive compared with small Gaussian masks.

After the smoothing stage, edge strength is found by taking the image gradient. Since there are edges in a variety of directions, four filters were used to detect vertical, horizontal and diagonal edges in the blurred image.

First derivative calculations (for horizontal and vertical direction) carried out in the edge detection algorithm returns values of \( G_x \) and \( G_y \). Direction and gradient can be computed from these values as follows:
\[ G = \sqrt{G_x^2 + G_y^2} \quad (A-2) \]
\[ \Theta = \arctan\left( \frac{G_y}{G_x} \right) \quad (A-3) \]

A.1.2 Symmetric Gaussian low pass filter

The Gaussian low pass filter is a 2-D convolution operator used to filter out high frequency details and noises and hence image is blurred. This filter is very similar to the mean filter in which all kernel components are the same. In this filter however a different kernel that represents the shape of a Gaussian hump.

The 1-D Gaussian distribution has the form:

\[ f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} \quad (A-4) \]

where \( \sigma \) is the standard deviation of the distribution. It is also assumed that the distribution has a mean of zero and distribution is centered on the line \( x=0 \).

The 2-D Gaussian distribution (i.e. circularly symmetric) similarly has the form:

\[ f(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (A-5) \]

The idea of Gaussian filtering is using its 2-D distribution to determine the convolution kernel by which new pixel intensity values will be computed. We need to produce a discrete approximation to the Gaussian function before performing the convolution.

Figure A-0-1 is an integer-valued convolution kernel that is obtained from a Gaussian distribution with a \( \sigma \) of 1.0.
Template matching is normally considered as a straightforward and relatively reliable method in character recognition. The first step in this method is normalising the character templates to a specific size (32 x 32 pixels in our implementation) and retaining it in the database. The character is firstly segmented from the number plate and after normalization is matched with all the characters in the database. The Hamming distance is the metric to measure the similarity between unknown character and normalized template.

This approach is shown as below:

\[
\frac{1}{273} \sum_{i=1}^{\text{nrows}} \sum_{j=1}^{\text{ncols}} \text{mismatch}_{i,j} \quad (A-6)
\]

Where

\[
\text{mismatch}_{i,j} = \begin{cases} 
1, & \text{if } \text{original}_{i,j} \neq \text{extracted}_{i,j} \\
0, & \text{if } \text{original}_{i,j} = \text{extracted}_{i,j}
\end{cases} \quad (A-7)
\]

Where \(\text{nrows}\) and \(\text{ncols}\) are the number of rows and columns (in our case \(\text{nrow}=\text{ncols}=32\)).

Using this technique, the character with the lowest mismatch value will be selected as the recognised character.

\[
\begin{array}{ccccc}
1 & 4 & 7 & 4 & 1 \\
4 & 16 & 26 & 16 & 4 \\
7 & 26 & 41 & 26 & 7 \\
4 & 16 & 26 & 16 & 4 \\
1 & 4 & 7 & 4 & 1 \\
\end{array}
\]

*Figure A-0-1  Discrete approximation to Gaussian function with \(\sigma=1.0\)*
A.3 THRESHOLDING TECHNIQUES

Most of the objects (e.g. license plates, face etc.) in surveillance videos are captured under heterogeneous lighting conditions. Shadowed regions can have very high light intensity or even sharp variation caused by reflected light. In this section, different techniques for thresholding number plates are explained as a specific example. Thresholding normally consists of two main parts. Rough thresholding is performed to identify a shadow type. In this stage, window sizes and locations are adjusted according to the shadow. In the next part of thresholding, numbers and characters are extracted after different contrast enhancement techniques.

Automatic number plate recognition is an essential element for developing an intelligent transportation system (ITS). License plate recognition (LPR) systems are composed of three main modules:

1- license plate detection
2- character segmentation
3- optical character recognition

Thresholding has been considered one of the most common pre-processing steps for character segmentation. Various thresholding methods have been used in different studies. Some of these methods select a single value to separate an image into object and background [251-253] or varying values (i.e. adaptive thresholds) are used in each block or window in some other studies [254-256]. Most of these methods have their own limitations, which can adversely affect other LPR modules. The Yang et al method relies on the assumption that the character region to license plate area is approximately fixed [251]. The threshold is adjusted according to this ratio by iterative thresholding. Having worked well in controlled areas, such methods will not give acceptable results under unrestricted illumination conditions. On the other hand, we have local thresholding in which the threshold is adjusted to a local intensity variation. Bernsen and Niblack introduced two well-known local thresholding techniques that are commonly used in the literature [254, 256]. For example, in Niblack’s method the threshold is computed from the mean and standard deviation of intensity values in a window and Brensen uses the maximum and minimum values in that window. These methods can be sensitive to local variation and their performance will be changed if we have significant local intensity variation. In few research studies, different thresholding methods were tested to find their performance under different lighting conditions. Bernsen’s method was selected as an effective method in one of these studies [257]. In some proposed approaches however, different combination of methods were used in order to improve the system performance [258, 259].

Many of these methods show good results in normal lighting conditions, but they show impaired performance where intensity changes drastically. Therefore, a robust thresholding method should be used to enable the system to segment the number plate and extract the alphanumeric characters.
efficiently when there is sharp intensity variation. In fact this thresholding technique can perform well in different lighting conditions specifically when there are shadowed regions, which would be the worst case scenario. It is not intended here to propose a new thresholding technique for number plates. However, there is the need to choose the very robust technique that can perform equally well in different lighting conditions. Many studies have been conducted on number plate localization and recognition, although not many of these use advanced thresholding techniques that can address complicated lighting condition issues. In this section existing methods are reviewed briefly and the selected method for the application used in our work is explained. The focus has been kept to proposed cases in which uneven illumination has been examined in their experiments. We have Niblack’s method in which local mean and local standard deviation are calculated \[ T(x, y) = m(x, y) + k \cdot s(x, y) \] (A-8)

In this equation \( m(x, y) \) is the average and \( s(x, y) \) is the standard deviation of luminance pixels in the predetermined window at \( (x, y) \). The size of the window is normally chosen by the user and should be small enough to enable it to reflect local details. Using too small a window size will also add to the system complexity and cannot suppress noise effects adequately either. Suitable values can be found by measuring character sizes in number plates. Variation of \( k \) in this equation can be used to adjust the foreground and background pixel contrast ratio. However this method does not work very well when there is uneven and sharp variation in the background. The modified Niblack version was proposed by Sanvola and Pietikainen in which the contribution of standard deviation was adaptively amplified with use of the local mean \[ T(x, y) = m(x, y) \left[ 1 - k \left( 1 - \frac{s(x, y)}{R} \right) \right] \] (A-9)

In this equation, \( R \) is the dynamic range of the standard deviation (e.g. \( R = 128 \) in 8-bit gray image). An alternative thresholding method based on boundary features was proposed by Wu et al. \[ \text{[258]} \]. In this method we can either select basic adaptive thresholding (BAT) or c-mean algorithm by using boundary characteristics. The c-mean algorithm is used when variances of two clusters are dissimilar and BAT is used in opposite case.

In this method the image is divided into \( M \times N \) macro-blocks and the intensity value of each pixel in \( (i,j)^{th} \) macro-block is indicated by \( f_{i,j}(x, y) \). There are also boundary pixels that are defined as follows:

\[ A_{i,j} = \{(x, y) | |f_{i,j}(x, y) - f_{i,j}(x-1, y)| \geq T_d \} \] (A-10)

Where \( T_d \) is a predefined threshold value in this equation.
In the next stage threshold is updated based on the average of the boundary pixels (i.e. $T_u$) and we can define the updated probability as follows:

\[
P_{u,0} = \frac{1}{M \times N} \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} \begin{cases} 
1 & \text{if } f_{i,j}(x, y) < T_u \\
0 & \text{otherwise}
\end{cases}
\]  

(A-11)

Such methods in which images are divided into macro-blocks are normally faster than pixel based methods [254, 256]. However, some unexpected ghost objects may appear, especially in uneven lighting images. Adaptive macro-block sizes based on data structure is also proposed to resolve these issues. Sizes are adaptively varied in relation to the Lorentz information measure (LIM) [260]. This measure helps determine whether or not the macro-block contains both the object and background.

There are many different factors that need to be considered in document thresholding such as illumination uniformity, contrast signal to noise ratio etc. Gatos et al [261] made a considerable effort to address and solve these issues by background surface thresholding. Despite being computationally expensive, this method has shown promising results in practice, particularly in challenging situations. Unfortunately no benchmark exists to compare all the above-mentioned proposed techniques’ performances under different conditions. Since we required a simple yet efficient technique in our pattern recognition system, the Sauvola method was exploited due to its comparatively simple to implement algorithm [255].

### A.4 AFFINE TRANSFORMATION

In many automatic pattern recognition systems, objects are subjected to geometric distortions. Such distortions occur when the distance of the camera to the object is changing, which can result in apparent object geometry alteration. An Affine transformation is a linear two dimensional geometric transformation which maps variables from input image into new variables in output image. This transformation applies a linear combination of translation, rotation, scaling and shearing operations (see Figure A-0-2). Different types of affine transformations have been widely used in computer vision application in the literature [262-266].

Arbitrary linear (or non-linear) transformations can be represented in a consistent format by matrices. This way we easily can concatenate different transformation by multiplying their matrices. 4x4 transformation matrices are very popular in 3D computer vision algorithms.
A.4.1 Translation:

Translation occurs when every point of the object is moved by the same amount in a given direction. In a simpler form, position of all points \((x, y)\) of objects will be changed to \((x + \Delta x, y + \Delta y)\). Another familiar term is a double reflection against two parallel axes that results in a total motion which is a translation.

A.4.2 Rotation:

To rotate an object by an angle of \(\theta\) (clockwise) about the origin below functions can be used:

\[
x' = x \cos \theta + y \sin \theta \quad \text{and} \quad y' = -x \sin \theta + y \cos \theta,
\]

or in matrix form:

\[
\begin{bmatrix}
  x' \\
  y'
\end{bmatrix} =
\begin{bmatrix}
  \cos \theta & \sin \theta \\
  -\sin \theta & \cos \theta
\end{bmatrix}
\begin{bmatrix}
  x \\
  y
\end{bmatrix}
\]

And for a counter clockwise rotation the functional form will be changed to:

\[
x' = x \cos \theta - y \sin \theta \quad \text{and} \quad y' = x \sin \theta + y \cos \theta
\]

Note that translation and rotation are among those transformations that do not distort the object shape.

A.4.3 Scaling:

To scale (enlarge or shrink) the object, we have \(x' = s_x x\) and \(y' = s_y y\) and in matrix form this will become:

\[
\begin{bmatrix}
  x' \\
  y'
\end{bmatrix} =
\begin{bmatrix}
  s_x & 0 \\
  0 & s_y
\end{bmatrix}
\begin{bmatrix}
  x \\
  y
\end{bmatrix}
\]

A.4.4 Reflection:

As previously mentioned linear transformations are normally represented by matrices and if \(T\) is a linear transformation mapping \(\mathbb{R}^n\) to \(\mathbb{R}^m\) and \(\vec{x}\) is a column vector with \(n\) components then \(T\) will be equal to: \(\bar{T}(\vec{x}) = A\vec{x}\), where \(A\) is an \(m \times n\) transformation matrix. To reflect an object about a line through the origin, assuming that \(\vec{l} = (l_x, l_y)\) is in direction of the line, \(A\) is going to be the transformation matrix:

\[
A = \frac{1}{\|\vec{l}\|^2}
\begin{bmatrix}
  l_x^2 - l_y^2 & 2l_y l_x \\
  2l_x l_y & l_y^2 - l_x^2
\end{bmatrix}
\]

\(\text{(A-12)}\)
In order to represent a two-vector \((x, y)\) as a three-vector \((x, y, 1)\) homogeneous coordinates will be used. Homogeneous coordinates are commonly used in projective geometry as opposed to cartesian coordinates used in Euclidean geometry. Homogeneous coordinates are widely used because they offer simpler transformation than their Cartesian counterparts and that is why they have been applied in a range of different applications including computer graphics and 3D computer vision. One main advantage of such coordinates is most affine transformations can be easily represented by a matrix. Using this system, object translation can be easily demonstrated with matrix multiplication. The functional form \(x' = x + t_x; y' = y + t_y\) becomes

\[
\begin{bmatrix}
x' \\
y' \\
1
\end{bmatrix} = \begin{bmatrix}
1 & 0 & t_x \\
0 & 1 & t_y \\
0 & 0 & 1
\end{bmatrix} \begin{bmatrix}
x \\
y \\
1
\end{bmatrix}
\]

and the clockwise rotation matrix from above becomes

\[
\begin{bmatrix}
x' \\
y' \\
1
\end{bmatrix} = \begin{bmatrix}
\cos \theta & \sin \theta & 0 \\
-\sin \theta & \cos \theta & 0 \\
0 & 0 & 1
\end{bmatrix} \begin{bmatrix}
x \\
y \\
1
\end{bmatrix}
\]
APPENDIX B

APPENDIX B MATLAB CODES

B.1 BIO-INSPIRED PRE-PROCESSING ALGORITHM

Some of the MATLAB codes used in this research work are considered in this appendix. The other codes and tools employed in this research have been provided on compact disk.

B.1.1 Main code

This is the main algorithm which contains Photoreceptor, LMC and EMD models.

System parameters have been slightly modified in different experiments. We can also exclude different system components if it is required.

```matlab
function FCr=MainCodeEx7
clc;close all;
% Dir='E:\Uni\video-footages\Character-Recog';
% Photoreceptor Parameters
% Phototime_1=4;
% Phototime_2=6;
% Phototime_3=8;
% LMC parameters
LMCtime_1 = 2.2;
LMCtime_2 = 12;
EMDtime=4;
%Thresholds
%Thr=6;
%thresh=[.03 .3];
% r1=1200;c1=1600;
%r1=520;c1=1280;
If=zeros(r1,c1,3);
BaseDir='E:\Uni\Matlab11\Vid-Sureveillance\JM\ffmpeg';
FrameRate=25;% Frame rate
i0=1;ie=20;
se = strel('rectangle',[15,25]);% kernel definition
se2 = strel('disk',16);% kernel definition
h = fspecial('disk',5);
filename1=strcat(BaseDir,'uncompress-dest17.avi');
Source=strcat(BaseDir,'Motion.avi');
source1=VideoReader(Source);
Obj1=VideoReader(filename1);
```

132
APPENDIX B

28. \( u=1; \)
29. \( \text{for } i=i_0:ie \)
30. \( \text{mov1.cdata = read(source1, } i); \)
31. \( \text{mov2.cdata = read(Obj1, } i); \)
32. \( \text{Im=mov1.cdata;} \)
33. \( \text{lc=mov2.cdata;} \)
34. \( \text{lc=im2double(lc(:,:,1));} \)
35. \( \text{inFrame=im2double(Im(:,:,1));} \)
36. \( \text{I0=inFrame;} \)
37. \( \text{if } i==i_0 \)
38. \( \text{f=1;} \)
39. \( \text{% initialization if its first frame} \)
40. \( \text{stage1LPFPrevState}=0; \text{stage2LPFPrevState}=0; \text{stage3LPFPrevState}=0; \)
41. \( \text{LMC1LPFPrevState}=0; \text{LMC2LPFPrevState}=0; \)
42. \( \text{EMDPrevState1}=0; \text{EMDShiftPrevState1}=0; \text{EMDShiftPrevState2}=0; \text{EMDShiftPrevState3}=0; \)
43. \( \text{End} \)
44. \( \text{% Photoreceptor Algorithm} \)
45. \( \text{[outPhoto,stage1LPFPrevState,stage2LPFPrevState,\ldots} \)
\( \text{\ldots stage3LPFPrevState]=Photoreceptor(inFrame,stage1LPFPrevState,\ldots} \)
\( \text{\ldots stage2LPFPrevState,stage3LPFPrevState,Phototime_1,Phototime_2,\ldots} \)
\( \text{\ldots Phototime_3,f);} \)
46. \( \text{outPhoto=inFrame;} \)
47. \( \text{% LMC Algorithm} \)
48. \( \text{[outLMC,LMC1LPFPrevState,LMC2LPFPrevState]=LMC(outPhoto,\ldots} \)
\( \text{\ldots LMC1LPFPrevState,LMC2LPFPrevState,\ldots} \)
49. \( \text{LMCtime_1,LMCtime_2,f);} \)
50. \( \text{outLMC(isnan(outLMC))=0;} \)
51. \( \text{outLMC=4*abs(outLMC);} \)
52. \( \text{% EMD Algorithm} \)
53. \( \text{[Result,EMDPrevState1,EMDShiftPrevState1,EMDShiftPrevState2,\ldots} \)
\( \text{\ldots EMDShiftPrevState3]=EMD(outLMC,EMDPrevState1,\ldots} \)
\( \text{\ldots EMDShiftPrevState3,EMDtime,f);} \)
54. \( \text{Ra=abs(Result);} \)
55. \( \text{% Normalization} \)
56. \( \text{Rn=(Ra-min(min(Ra)))/(max(max(Ra))-min(min(Ra)))}; \)
57. \( \text{Rn=Rn.*(Rn>.08);} \)
58. \( \text{%**** Cleaning (Erosion and Dilation ))********} \)
59. \( \text{R = edge(Rn,'Canny',thresh);} \)
60. \( \text{R=imclose(R,se2); } \)
61. \( \text{R=medfilt2(R,[40 40]); } \)
62. \( \text{R = imdilate(R,se);} \)
63. \( \text{R=I0.*R;} \)
64. \( \text{R = edge(R,'Canny',thresh);} \)
65. \( \text{R=imclose(R,se2);} \)
B.1.2 Photoreceptor

This is the photoreceptor source code which is included in the main Algorithm. Its parameters however are selected in the main algorithm.

```matlab
function [outFrame,stage1LPFPrevState,stage2LPFPrevState,stage3LPFPrevState]= ...Photoreceptor(inFrame,stage1LPFPrevState,...
stage2LPFPrevState,stage3LPFPrevState,Phototime_1,Phototime_2,Phototime_3,Ini)

time_constant_1=Phototime_1;
time_constant_2=Phototime_2;
time_constant_3=Phototime_3;
dark_gain =60;
mid_dark_gain =0.1 ;
naka_rushton_constant =.6;
% divisor_limit = .1;
% exp_divisor_limit =100;
%Initialization
if (time_constant_1<1)
time_constant_1 = 1;
end;
if (time_constant_2<1)
time_constant_2 = 1;
end;
if (time_constant_3<1)
time_constant_3 = 1;
end;
%Initialization (This is carried out only in the first iteration )
if (Ini == 1)
stage1LPFPrevState= inFrame;
stage1Output= dark_gain*(1-stage1LPFPrevState./(stage1LPFPrevState +
mid_dark_gain));
stage1Output = stage1Output.*inFrame;
stage2LPFPrevState = sqrt(inFrame)./time_constant_2;
divisor = stage2LPFPrevState;
%divisor=divisor_limit.*(divisor<divisor_limit)+divisor.*(divisor>=divisor_limit);
stage2Output= stage1Output./divisor;
stage3Output = 0.7*log((stage2Output + exp(1)))*0.5 - 0.7*log(0.5*exp(1));
stage3LPFPrevState = stage3Output./time_constant_3;
expDivisor = 3.67882*(2.^stage3LPFPrevState);
stage3Output = stage2Output./expDivisor;
outFrame = stage3Output./(stage3Output + naka_rushton_constant);
else
stage1LPFPrevState = (inFrame - stage1LPFPrevState)./time_constant_1 +
...stage1LPFPrevState;
stage1Output = dark_gain*(1-stage1LPFPrevState./(stage1LPFPrevState +
```
APPENDIX B

37. …mid_dark_gain));
38. stage1Output = (stage1Output).*inFrame;
39. stage2LPFPrevState = (stage1Output - stage2LPFPrevState)/time_constant_2 + 
    …stage2LPFPrevState;
40. divisor = stage2LPFPrevState;
41. % divisor = divisor_limit.*(divisor<divisor_limit)+divisor.*(divisor>=divisor_limit);
42. stage2Output = stage1Output/divisor;
43. stage3LPFPrevState = (stage2Output - stage3LPFPrevState)/time_constant_3 + 
    …stage3LPFPrevState;
44. expDivisor = 3.67882.*(2.^stage3LPFPrevState);
    % expDivisor = exp_divisor_limit.*(expDivisor<exp_divisor_limit)+expDivisor.*(exp…Divisor>=exp_divisor_limit);
45. stage3Output = stage2Output/expDivisor;
46. outFrame = stage3Output/(stage3Output + naka_rushton_constant);
47. end

B.1.3 LMC

This is the LMC source code which is included in the main Algorithm. Some parameters however are 
set in the main algorithm and the other parameters are internally set in this function.

1. function [outFrame, LMC1LPFPrevState, LMC2LPFPrevState] = …
    …LMC(inFrame, LMC1LPFPrevState, LMC2LPFPrevState, t_constant_1, …
    …t_constant_2, f)
2. K1 = 0.4;
3. % K1 = 0.21;
4. K2 = 0.72;
5. K3 = 2;
6. h = [0,0,1,0,0,0,1,2,1,0,1,2,4,2,1,0,1,2,1,0,0,0,1,0,0]/20; % Gaussian Kernel
7. % High Pass filtering by Low Pass filtering Subtraction
8. HPFFrame = inFrame - imfilter(inFrame, h,'replicate');
9. if (f == 1)
10. inFrame = inFrame - K1;
11. LMC1LPFPrevState = sqrt(inFrame) / t_constant_1;
12. LMC2LPFPrevState = sqrt(HPFFrame) / t_constant_2;
13. LMC1Output = inFrame - K2*LMC1LPFPrevState;
14. outFrame = tanh(K3*(LMC1Output - LMC2LPFPrevState));
15. else
16. inFrame = inFrame - K1;
17. LMC1LPFPrevState = (inFrame - LMC1LPFPrevState)/t_constant_1 + 
    …LMC1LPFPrevState;
18. LMC1Output = inFrame - (K2*LMC1LPFPrevState);
19. LMC2LPFPrevState = (HPFFrame - LMC2LPFPrevState)/t_constant_2 + 
    …LMC2LPFPrevState;
20. outFrame = tanh(K3*(LMC1Output - LMC2LPFPrevState));
B.1.4 EMD

This is the EMD source code which is included in the main Algorithm. EMD parameters are similarly defined in the main algorithm.

```matlab
inFrame=I;
[r c]=size(inFrame);
Shift1=I;
Shift2=I;
Shift3=I;
for i=2:r-1
    for j=2:c-1
        Shift1(i,j)=I(i-1,j+1);
        Shift2(i,j)=I(i,j+1);
        Shift3(i,j)=I(i+1,j+1);
    end
end
if (f == 1)
    EMDPrevState1= inFrame;
    EMDShiftPrevState1 = Shift1;
    EMDShiftPrevState2= Shift2;
    EMDShiftPrevState3 = Shift3;
    TempUp1=EMDPrevState1.*Shift1;
    TempUp1=tanh(TempUp1);
    TempDown1=EMDShiftPrevState1.*inFrame;
    TempDown1=tanh(TempDown1);
    % ************************************************
    %  ************************************************
    TempUp2=EMDPrevState1.*Shift2;
    TempUp2=tanh(TempUp2);
    TempDown2=EMDShiftPrevState2.*inFrame;
    TempDown2=tanh(TempDown2);
    % ************************************************
    %  ************************************************
    TempUp3=EMDPrevState1.*Shift3;
    TempUp3=tanh(TempUp3);
    TempDown3=EMDShiftPrevState3.*inFrame;
    TempDown3=tanh(TempDown3);
    % ************************************************
    %  ************************************************
end
```

% end
APPENDIX B

37. Result=.7*(TempUp1-TempDown1)+(TempUp2-TempDown2)+.7*(TempUp3-
38. ....TempDown3);
39. else
40. % % //PERFORM THE PROCESSING
41. % % //low pass filtering
42. % % //output becomes previous state for next frame LPF
43. EMDPrevState1 = (inFrame - EMDPrevState1)/t_constant + EMDPrevState1;
44. EMDShiftPrevState1 = (Shift1 - EMDShiftPrevState1)/t_constant +
...EMDShiftPrevState1;
45. TempUp1=EMDPrevState1.*Shift1;
46. TempUp1=tanh(TempUp1);
47. TempDown1=EMDShiftPrevState1.*inFrame;
48. TempDown1=tanh(TempDown1);
49. % *********************************************
50. **********************************************
51. EMDShiftPrevState2 = (Shift2 - EMDShiftPrevState2)/t_constant +
...EMDShiftPrevState2;
52. TempUp2=EMDPrevState1.*Shift2;
53. TempUp2=tanh(TempUp2);
54. TempDown2=EMDShiftPrevState2.*inFrame;
55. TempDown2=tanh(TempDown2);
56. % *********************************************
57. **********************************************
58. % % //EMDPrevState3 = EMDPrevState1;
59. EMDShiftPrevState3 = (Shift3 - EMDShiftPrevState3)/t_constant +
...EMDShiftPrevState3;
60. TempUp3=EMDPrevState1.*Shift3;
61. TempUp3=tanh(TempUp3);
62. TempDown3=EMDShiftPrevState3.*inFrame;
63. TempDown3=tanh(TempDown3);
64. Result=.7*(TempUp1-TempDown1)+(TempUp2-TempDown2)+.7*(TempUp3-
65. ....TempDown3);
66. end
Bibliography


[104] H. T. Sencar and N. Memon, Digital image forensics: There is more to a picture than meets the eye: Springer, 2012.
BIBLIOGRAPHY


BIBLIOGRAPHY


BIBLIOGRAPHY


BIBLIOGRAPHY


