APPLICATION OF FOOTSTEP SOUND AND LAB COLOUR SPACE IN MOVING OBJECT DETECTION AND IMAGE QUALITY MEASUREMENT USING OPPOSITE COLOUR PAIRS

by

Aditya Roshan

M Tech in Civil Engineering, Indian Institute of Technology Kanpur, 2010

A Dissertation Submitted in Partial Fulfilment of the Requirements for the Degree of

Doctor of Philosophy

in the Graduate Academic Unit of Geodesy and Geomatics Engineering

Supervisor: Yun Zhang, PhD, Geodesy and Geomatics Engineering

Examining Board: David Coleman, PhD, Geodesy and Geomatics Engineering
Emmanuel Stefanakis, PhD, Geodesy and Geomatics Engineering
Dr. Howard Li, PhD, Electrical and Computer Engineering

External Examiner: Dr. Naser El-Sheimy, PhD, Dept. of Geomatics Engineering, University of Calgary

This dissertation is accepted by the Dean of Graduate Studies

THE UNIVERSITY OF NEW BRUNSWICK

February, 2019

© Aditya Roshan, 2019
Abstract

This PhD dissertation is focused on two of the major tasks of an Atlantic Innovation Fund (AIF) sponsored “Triple-sensitive Camera Project”. The first task focuses on the improvement of moving object detection techniques, and second on the evaluation of camera image quality. Cameras are widely used in security, surveillance, site monitoring, traffic, military, robotics, and other applications, where detection of moving objects is critical and important. Information about image quality is essential in moving object detection. Therefore, detection of moving objects and quality evaluation of camera images are two of the critical and challenging tasks of the AIF Triple-sensitive Camera Project.

In moving object detection, frame-based and background-based are two major techniques that use a video as a data source. Frame-based techniques use two or more consecutive image frames to detect moving objects, but they only detect the boundaries of moving objects. Background-based techniques use a static background that needs to be updated in order to compensate for light change in a camera scene. Many background modelling techniques involving complex models are available which make the entire procedure very sophisticated and time consuming. In addition, moving object detection techniques need to find a threshold to extract a moving object. Different thresholding methodologies generate varying threshold values which also affect the results of moving object detection. When it comes to quality evaluation of colour images, existing Full-Reference methods need a perfect colour image as reference and No-Reference methods use a gray image generated from the colour image to compute image quality. However, it is very challenging to find a perfect reference colour image. When a colour image is converted to a grey image for image quality evaluation, neither colour information nor
human colour perception is available for evaluation. As a result, different methods give varying quality outputs of an image and it becomes very challenging to evaluate the quality of colour images based on human vision.

In this research, a single moving object detection using frame differencing technique is improved using footstep sound which is produced by the moving object present in camera scene, and background subtraction technique is improved by using opposite colour pairs of Lab colour space and implementing spatial correlation based thresholding techniques. Novel thresholding methodologies consider spatial distribution of pixels in addition to the statistical distribution used by existing methods. Out of four videos captured under different scene conditions used to measure improvements, a specified frame differencing technique shows an improvement of 52% in object detection rate when footstep sound is considered. Other frame-based techniques using Optical flow and Wavelet transform such are also improved by incorporating footstep sound. The background subtraction technique produces better outputs in terms of completeness of a moving object when opposite colour pairs are used with thresholding using spatial autocorrelation techniques. The developed technique outperformed background subtraction techniques with most commonly used thresholding methodologies. For image quality evaluation, a new “No-Reference” image quality measurement technique is developed which evaluates quantitative image quality score as it is evaluated by human eyes. The SCORPIQ technique developed in this research is independent of a reference image, image statistics, and image distortions. Colour segments of an image are spatially analysed using the colour information available in Lab colour space. Quality scores from SCORPIQ technique using LIVE image database yield distinguished results as compared to quality scores from existing methods which give
similar results for visually different images. Compared to visual quality scores available with LIVE database, the quality scores from SCORPIQ technique are 3 times more distinguishable. SCORPIQ give 4 to 20 times distinguishable results compared to statistics based results which also does not follow the quality scores as evaluated by human eyes.
Acknowledgement

Many individuals have helped me to successfully achieve the objectives of this research and I would like to express my gratitude to them.

First, I would like to thank my supervisor Dr. Yun Zhang for guiding me through my research work and providing valuable advice at all times. He taught me valuable lessons in research by patiently reviewing my papers, thesis, and presentations. Furthermore, he provided me opportunities for attending conferences and other academic activities.

I am very grateful to Dr. David J. Coleman and Dr. Emmanuel Stefanakis for their help in my major and minor examinations, and also for reviewing my proposal, and dissertation. I am also thankful to Dr. Marcelo Santos for his encouragement throughout the program and Dr. Monica Wachowicz for giving lessons about report writing.

I acknowledge Atlantic Canada Opportunities Agency and New Brunswick Innovation Fund programs for providing financial support for this research.

Many thanks to staff at the Department of Geodesy and Geomatics Engineering, especially David Fraser, Sylvia Whitaker, and Lorry Hunt for their kind support. I would also like to thank Sina Adham Khiabani, Fatemeh Fathollahi, Mark Doucette and Dr. Paul Hink for their guidance with the Triple Sensitive Camera Project. Many thanks to the UNB Writing Centre for their advice in writing and reviewing papers and reports.

Finally, I would like to thank my wife Heena for helping me through the entire process by not only providing me with unconditional love and understanding, but also supporting me by giving critical reviews on my research work.
# Table of Contents

Abstract ................................................................................................................................. ii

Acknowledgement ............................................................................................................... v

Table of Contents ............................................................................................................... vi

List of Tables ....................................................................................................................... xi

List of Figures ..................................................................................................................... xii

List of Symbols .................................................................................................................. xvi

Chapter 1 : Introduction ...................................................................................................... 1

1.1 Dissertation Structure................................................................................................. 2

1.2 Background .................................................................................................................. 3

1.3 Selection of Research Topics ...................................................................................... 3

1.4 Review of Existing Solutions ...................................................................................... 6

1.4.1 Frame-based moving object detection ................................................................. 7

1.4.2 Background-based moving object detection ......................................................... 8

1.4.3 No-Reference image quality measurement .......................................................... 10

1.5 Problem statement ....................................................................................................... 11

1.6 Research Objectives ................................................................................................... 13

1.7 Data and Metrics ......................................................................................................... 13

1.8 Overview of each chapter .......................................................................................... 15

1.9 References .................................................................................................................. 17

Chapter 2 : Using Mel-Frequency Audio Features from Footstep Sound and Spatial
Segmentation Techniques to Improve Frame-based Moving Object Detection .......... 22
2.1 Introduction .................................................................................................................. 23
2.2 Literature Review ......................................................................................................... 26
2.3 Methodology .................................................................................................................. 29
2.3.1 Moving Object Detection .......................................................................................... 29
2.3.2 Audio Processing for Footstep Sound Detection ...................................................... 31
2.3.3 Video Frames Counting .............................................................................................. 33
2.3.4 Missing, Partially and Noisy Detected Object .............................................................. 33
2.4 Results and Discussion .................................................................................................. 36
2.4.1 Camera System ........................................................................................................... 36
2.4.2 Camera Scene ............................................................................................................. 37
2.4.3 Manual and Otsu Thresholding for Frame Differencing ............................................. 37
2.4.4 Footstep Sound and Frame Counting ......................................................................... 39
2.4.5 Modelling Moving Object in Frames with Partially Detected, Missing Object and Noisy Output ........................................................................................................... 41
2.4.6 Moving Object Detection .......................................................................................... 43
2.5 Conclusions .................................................................................................................... 46
2.6 Acknowledgements ........................................................................................................ 47
2.7 References ..................................................................................................................... 48

Chapter 3: Lab Colour Space and Spatial Correlation based Moving Object Detection
............................................................................................................................................ 51

vii
3.1 Introduction.................................................................................................................. 52
3.2 Review of Existing Image Thresholding Methods .................................................. 57
  3.2.1 Otsu Threshold Method......................................................................................... 57
  3.2.2 Adaptive Image Threshold Method....................................................................... 58
  3.2.3 Iterative Gaussian Clustering Method................................................................. 59
  3.2.4 Histogram-Based Thresholding Method............................................................... 60
  3.2.5 Moving Average Histogram Thresholding ......................................................... 61
  3.2.6 Histogram Entropy-Based Image Thresholding .................................................. 61
  3.2.7 Minimum Error Thresholding................................................................................ 62
3.3 Methodology................................................................................................................ 63
  3.3.1 RGB Colour Space to Lab Colour Space Conversion........................................... 63
  3.3.2 Background and Foreground Subtraction............................................................. 65
  3.3.3 Spatial Autocorrelation......................................................................................... 65
3.4 Results and Discussion ............................................................................................... 71
  3.4.1 Image Dataset for Moving Object Detection....................................................... 71
  3.4.2 Moving Object....................................................................................................... 72
3.5 Conclusions.................................................................................................................. 77
3.6 Acknowledgement ...................................................................................................... 78
3.7 References.................................................................................................................... 78

Chapter 4 : SCORPIQ – A Referenceless Perceptual Image Quality Measurement
Technique for Colour Images .......................................................................................... 83
4.1 Introduction ........................................................................................................... 83
4.2 Literature Review ................................................................................................ 87
4.3 Methodology ......................................................................................................... 90
  4.3.1 Image Conversion and Processing ................................................................... 90
  4.3.2 Image Quality Measurement .......................................................................... 92
4.4 Results and Discussions ....................................................................................... 98
  4.4.1 Quality of images with varying colours at same pixel location ................. 99
  4.4.2 Quality of images with similar colours at different locations ................. 101
  4.4.3 Colour and Spatially Variant Image Quality ............................................. 102
4.5 Conclusions .......................................................................................................... 104
4.6 Acknowledgements .............................................................................................. 106
4.7 References ............................................................................................................. 106

Chapter 5 : Summary and Conclusions ..................................................................... 110

5.1 Summary of Research ......................................................................................... 110
  5.1.1 Summary of Chapter 2 .................................................................................. 110
  5.1.2 Summary of Chapter 3 .................................................................................. 111
  5.1.3 Summary of Chapter 4 .................................................................................. 112
5.2 Research Achievements ....................................................................................... 113
  1.1.1 Improved Frame Differencing moving object detection technique .......... 114
  1.1.2 Improved Background Subtraction moving object detection technique 114
1.1.3 Development of a new No-Reference image quality measurement technique 115

5.3 Suggestions for future work................................................................. 115
List of Tables

Table 1.1. Data, instrument and metrics used for moving object detection and quality assessment

Table 2.1 Camera and audio details

Table 2.2 Frame counts in videos during footstep sound and frame differencing detection analysis

Table 2.3 Improvements to the frame-based moving object detection techniques when footstep sound is detected

Table 3.1. Threshold selection criteria using spatial correlation coefficients

Table 3.2. Comparison of moving object detection results using six different techniques of image thresholding

Table 4.1. Quality score of images presented in Figure. 4.8

Table 4.2. Quality score of images in Figure. 4.9

Table 4.3. Quality scores of images in Figure. 4.10
List of Figures

Figure 1.1 Organization of thesis document ................................................................. 2

Figure 2.1. Process showing the improvement of Frame Differencing (FD) moving object
detection technique using footstep sound recorded with video recording system .......... 25

Figure 2.2. Audio processing and video frame counting for each peak location in input
sound .................................................................................................................................. 31

Figure 2.3. Camera scenes in captured videos .................................................................... 37

Figure 2.4. Frame differencing moving object detection results using manual and Otsu
thresholding. Row 1: manual threshold and Row 2: Otsu threshold (Frames from left to
right: Video 1, frame 30; Video 1, frame 184; Video 2, frame 30 and Video 2, frame 254).
Row 3: manual threshold and Row 4: Otsu threshold (Frames from left to right: Video 3,
frame 147; Video 3, frame 482; Video 4, frame 132 and Video 4, frame 214) ................. 38

Figure 2.5. Video 1 footstep sound waveform and detection of peak locations
 corresponding to footstep sound using first MFCC coefficients ...................................... 39

Figure 2.6. Frame differencing output frame analysis based on fully detected object falling
within the duration of detected footstep sound ................................................................. 40

Figure 2.7. Moving object detection and modelling of missing/partially/noisy moving
object using neighbouring frames. Three consecutive frames within the presence of
footstep sound are shown. Object in the middle frame is modelled using previous and next
frames. Rows 1 & 2: Video 1, Frames 166, 167 and 168. Rows 3 & 4: Video 2, Frames
348, 349 and 350. Rows 5 & 6: Video 3, Frames 359, 360 and 361. Rows 7 & 8: Video 4,
Frames 171, 172 and 173 .................................................................................................. 42
Figure 2.8. Noisy frames and noise removal using spatial segmentation. Row 1: Frame 187 of Video 1, Row 2: Frame 3 of Video 2, Row 3: Frame 570 of Video 3 and Row 4: Frame 314 of Video 4. Column 1: Frame differencing output frame, Column 2: Spatially segmented pixels, Column 3: Frames after removing noise, and Column 4: Object modelled after noise removal...

Figure 2.9. Object modelled as ellipse to frames with missing/partially detected object and noisy outputs. Row 1: Video 1, Frames 27, 74, 199, 234. Row 2: Video 2, Frames 260, 308, 316, 385. Row 3: Video 3, Frames 250, 270, 419, 540. Row 4: Video 4, Frames 21, 133, 150, 332...

Figure 2.10. Output frames of moving object detection using optical flow and wavelet transform based moving object detection techniques. Sample video frames within duration of footstep sound from Videos 1 to 4. Row 1: video output frames using optical flow and Row 2: video output frames using wavelet transform...

Figure 3.1. Iterative Gaussian Clustering Algorithm...

Figure 3.2. Spatial correlation-based image thresholding algorithm...

Figure 3.3. Spatial correlation (Moran’s Index) calculation between two image bands...

Figure 3.4. Indoor and outdoor camera scenes...

Figure 3.5. Different moving objects in different camera scenes...

Figure 3.6. Moving object detection in indoor scene and medium light condition...

Figure 3.7. Moving object detection in indoor scene and high light condition...

Figure 3.8. Moving object detection in outdoor scene and daytime light condition...

Figure 3.9. Detailed detected object in indoor and outdoor scenes...
Figure 4.1. Three images captured using one camera at one location. Scene condition is different in each image.

Figure 4.2. Lab space color wheel (a) Lab color space coordinate system (b) Location of primary RGB colors (Red [255, 0, 0], Green [0, 255, 0], Blue [0, 0, 255] and Yellow [255, 255, 0]) on color wheel (after Inc., 2018)

Figure 4.3. A camera produced image as seen by human eyes. (a) RGB image where objects are constructed by spatially arranging RGB colours (Sheikh, et al., 2005), (b) Representation of colours as seen in (a) in three dimensional Lab color space (x: Red (+a) – Green (-a), y: Yellow (+b) – Blue (-b) and z: Black (L = 0) – White (L = 100)

Figure 4.4. Steps for image quality measurement in Lab color space

Figure 4.5. Image segmentation (a) Input image (b) Sample region (c) segmented pixels (d) Segment boundary

Figure 4.6. Segment neighborhood and sub-segment analysis, consisting of the following steps (a) Find the angular extent of all segments, (b) Eliminate far and overlapping angular extent of neighborhood segments, (c) Calculate boundaries of segment under consideration falling within each angular extent and (d) Represent each part of segment as sub-segment.

Figure 4.7. Color segments and their boundaries (a) Segment boundaries detected in Lab color space (b) Segment boundaries after removing unwanted small segment and eliminating overlapping areas

Figure 4.8. Color chart images (RGB) for image quality measurements. Original image (a) is modified in Lab color space. Lightness is compressed to 0.7 and 0.3 of original lightness and colors are stretched to 2 and 3 times of original values respectively.
Figure 4.9. Image generated after changing spatial locations of color segments on original color chart (Figure 4.8(a)) ................................................................. 101

Figure 4.10. Non-distorted image set from LIVE database (Sheikh, et al., 2005)) ....... 103
# List of Symbols

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACOA:</td>
<td>Atlantic Canada Opportunities Agency</td>
</tr>
<tr>
<td>AIF:</td>
<td>Atlantic Innovation Fund</td>
</tr>
<tr>
<td>BRISQUE:</td>
<td>Blind/Referenceless Image Spatial Quality Evaluator</td>
</tr>
<tr>
<td>CCD:</td>
<td>Charge Coupled Device</td>
</tr>
<tr>
<td>CIELAB:</td>
<td>International Commission on Illumination Lab</td>
</tr>
<tr>
<td>CMOS:</td>
<td>Complementary Metal Oxide Semiconductor</td>
</tr>
<tr>
<td>CNN:</td>
<td>Conventional Neural Network</td>
</tr>
<tr>
<td>DCT:</td>
<td>Discrete Cosine Transform</td>
</tr>
<tr>
<td>DFT:</td>
<td>Discrete Fourier Transform</td>
</tr>
<tr>
<td>DIIVINE:</td>
<td>Distortion Identification-based Image Verity and Integrity Evaluation</td>
</tr>
<tr>
<td>DWT:</td>
<td>Discrete Wavelet Transform</td>
</tr>
<tr>
<td>FD:</td>
<td>Frame Differencing</td>
</tr>
<tr>
<td>FPS:</td>
<td>Frames Per Second</td>
</tr>
<tr>
<td>FR:</td>
<td>Full Reference</td>
</tr>
<tr>
<td>FSIM:</td>
<td>Feature Similarity Index Measure</td>
</tr>
<tr>
<td>GMM:</td>
<td>Gaussian Mixture Model</td>
</tr>
<tr>
<td>HVS:</td>
<td>Human Visual System</td>
</tr>
<tr>
<td>Hz:</td>
<td>Hertz</td>
</tr>
<tr>
<td>JND:</td>
<td>Just Noticeable Difference</td>
</tr>
<tr>
<td>Lab:</td>
<td>L for Lightness, a and b for red-green and blue-yellow colour pairs</td>
</tr>
<tr>
<td>LISA:</td>
<td>Local Indicators of Spatial Association</td>
</tr>
<tr>
<td>MFCC:</td>
<td>Mel Frequency Cepstral Coefficients</td>
</tr>
</tbody>
</table>
MP: Mega Pixels
MSE: Mean Square Error
NIIRS: National Imagery Interpretability Rating Scale
NR: No Reference
NSS: Natural Scene Statistics
PSNR: Peak Signal to Noise Ratio
QAC: Quality Aware Clustering
RGB: Red, Green, Blue
RMSE: Root Mean Square Error
RR: Reduced Reference
SCORPIQ: Spatial COlour Referenceless Perceptual Image Quality
SNR: Signal to Noise Ratio
SSIM: Structural SIMilarity
TCP: Triple-sensitive Camera Project
Th: Threshold
UNB: University of New Brunswick
VIF: Visual Information Fidelity
Chapter 1: Introduction

This PhD dissertation highlights existing problems and presents their solutions in the fields of moving object detection utilizing video stream from a camera and colour image quality evaluation of images obtained from the camera. These are two of the major tasks of an Atlantic Innovation Fund (AIF) sponsored “Triple-sensitive Camera Project”. This is a paper-based dissertation presented on the basis of findings mentioned in the documents below:


1.1 Dissertation Structure

This research document consists of 5 chapters. Three peer-reviewed journal papers, published or under review, are incorporated in the dissertation. Chapter 1 provides an overview of the research, Chapters 2 to 4 present the three journal papers, and Chapter 5 describes the research summary and conclusions.

Figure 1.1 illustrates the organization of this dissertation.

Figure 1.1 Organization of thesis document
1.2 Background

Images and videos captured by cameras are widely used in extracting valuable information in many applications such as site monitoring, face detection, remote sensing, surveillance, military, traffic, crime, robotics, medicine, navigation, and communications (Kumar, et al., 2016; Kamate & Yilmazer, 2015). Different applications require different payloads and the camera cost increases with the demands of cameras of higher sensitivities (or higher signal-to-noise ratio). The Atlantic Innovation Fund (AIF) sponsored Triple-sensitive Camera Project (TCP) was started in 2012 to develop a prototype for a super camera which would be more sensitive than other existing cameras, and be able to detect moving objects present in the camera scene. Moving object detection, implemented as a module is an integral part of the AIF Triple-Sensitive Camera Project. Image quality measurement is needed to demonstrate and prove the improvement of camera sensitivity which is visible to human eyes. Existing No Reference methods of image quality measurement are based of image statistics (such as Root Mean Square Error, Peak Signal to Noise Ratio, Fourier Transform, Discrete Cosine Transform) which does give quality scores of images as perceived by human eyes.

1.3 Selection of Research Topics

The Triple Sensitive Camera project needed an investigation on techniques suitable for moving object detection and required to evaluate image (and video) quality according to human eye perception. Both the tasks were challenging for AIF Triple-Sensitive Camera Project.
In the area of moving object detection, the existing techniques can be divided into two major groups: frame-based and background-based techniques. Since Frame-based techniques does not involve statistical modeling or neural network learnings, the processing of difference between two or more consecutive image frames is less complex than background-based techniques where statistical modeling and learnings are used to update the background frame. Frame-based techniques are computationally fast but only detect the boundary of the moving object whereas background-based techniques detect the complete moving object but need to update background for any scene change.

Results of moving object detection are affected by various scene condition properties such as light change in camera scene, poor contrast between moving object and background, camera sensor noise, shadows, dynamic background, clutter and camouflage. Consequently, noisy pixels are detected as moving object which are minimized using image thresholding methods. Image thresholding plays a critical role in moving object detection and is used with moving object detection techniques to extract the moving object. Appropriate selection of threshold value is another challenge which is affected by changes in the scene condition.

Moving object detection using frame differencing is used sparingly due to poor outputs despite it being computationally inexpensive among frame-based techniques (Roshan & Zhang, 2014). Frame differencing techniques only detect the boundary of a moving object, due to which an object is either detected partially or fails to detect completely. This makes it difficult to determine the presence of a moving object in camera scene. Background subtraction is the most commonly used background-based technique but continuous scene changes also need a frequent background update. Different background subtraction
techniques are developed utilizing various statistical models to update the background. Complex background models increase algorithm complexity which reduces the computational speed. As a result, demand for efficient processing systems increases. There also exist training based moving object detection techniques which require extensive training and a dedicated processing unit (or onboard integration of processing algorithms) for the detection of moving objects.

Camera technology and image processing algorithms are developing with time. Different cameras produce images for various applications such as remote sensing, medical, photography, security, transportation, and navigation. Based on the application, image quality measurement and image utilization methods can vary. Image quality score is a numeric representation of images as seen by human eyes or a mathematical model including image quality parameters. There exist quantitative (or objective) and qualitative (or subjective) image quality evaluation techniques. These techniques can be divided into three groups: Full-Reference, Reduced-Reference and No-Reference. Full-Reference techniques require reference colour image to evaluate image quality whereas Reduced-Reference techniques require reference image only for the purpose of building an evaluation model. Finally, No-Reference image quality measurement techniques do not require any reference image as they use image statistics to evaluate image quality.

No-Reference image quality measurement is useful in the sense that it is very challenging to find a perfect reference colour image to compute image quality using Full or Reduced-Reference techniques. However, existing No-Reference image quality measurement techniques compute image quality using image statistics from gray images which do not consider the spatial distribution of objects in an image. When a coloured
image is converted to a gray image, it loses the colour information. Also, the quality computed using a gray image does not consider the colour perception model of human vision. It is possible that two or more differently appearing images look the same when converted to gray and possess the same image statistics.

The primary objectives of this research are: 1) to evaluate existing frame-based and background-based moving object detection techniques and improve them for better object detection rates under different scene conditions, and 2) to evaluate No-Reference image quality measurement techniques and develop a No-Reference image quality measurement technique which can determine colour image quality as it is perceived by human eyes without the need for a reference image.

1.4 Review of Existing Solutions

Different camera manufacturing companies use different camera sensor technologies (CMOS and CCD sensors) and processing algorithms (Malams, et al., 2003). With technological and processing differences, images/videos of a scene are captured differently. Noise profiles and colours are also captured differently by different sensors (Carlson, 2002; Hain, et al., 2007). Due to these differences, moving object detection techniques give different results. In the literature, there are many moving object detection and image quality measurement techniques. Both can be classified into two groups based on the basic phenomenon used.

Moving object detection:
• Group 1 moving object detection techniques are frame-based and use frame differencing as the basic operation where two or more consecutive frames are subtracted to detect the moving object (Dedeoglu, 2004).

• Group 2 moving object detection techniques are background-based and use background subtraction as the basic operation where a static frame without any object is subtracted from each incoming frame (McIvor, 2000).

Image quality evaluation:

• Group 1 image quality measurement techniques compare a reference image to an image to generate a quality score with respect to the reference image (Li, et al., 2016).

• Group 2 image quality measurement techniques do not require a reference image -- a standalone quality score is obtained using image statistics and image pixel data (Gabarda & Cristóbal, 2007).

1.4.1 Frame-based moving object detection

Frame differencing is the most commonly used frame-based moving object detection technique. It uses two successive frames (at time $t$ and $t-1$) from a video sequence and subtracts them pixel by pixel. This technique only detects the boundary of moving objects but is highly adaptive to dynamic scene changes (Dedeoglu, 2004). Two frame differences from three consecutive frames are combined using the logical AND operation in double frame differencing technique (Kameda & Minoh, 1996). This technique addresses certain loopholes in the frame differencing technique by detecting more complete boundaries of the moving objects. Unwanted noise is removed from frame differences by using a
threshold value. The pixels with higher difference value (greater than threshold) are masked as moving object pixels. Being computationally less complex and faster in processing, these techniques have not been used much mainly due to poor quality of outputs and limited functionality of only detecting boundaries of moving objects.

Wavelet Transform and Optical flow are also used with frame-based moving object detection techniques. Wavelet coefficients computed out of double difference using Wavelet Transform (Kameda & Minoh, 1996; Antic, et al., 2009; Baradarani & Wu, 2010) and Optical Flow velocity generated from the relative motion between two consecutive frames are subjected to threshold values in order to detect the moving object. Optical Flow velocity gives important information about the spatial arrangement of objects in a scene and rate of change of their arrangement patterns (Shafie, et al., 2009; Horn & Schunck, 1981). Both Wavelet- and Optical Flow-based moving object detection techniques increase the complexity without offering any major improvements to the detection results. Frame-based moving object detection techniques are sensitive to noise and fail to detect moving objects if they are moving slowly or if they stop while moving. Poor contrast between object and its background, camouflage, clutter, shadows, and application of multiple light sources further affect the outputs from these techniques.

1.4.2 Background-based moving object detection

Background subtraction techniques are the most popular moving object detection techniques (McIvor, 2000) where moving objects are detected by computing the difference between the background and incoming frame. This difference is subjected to a threshold value. Different background-based techniques use background modeling to improve
detection outputs (Toyama, et al., 1999). Different ways of modeling the background have been tried (Elhabian, et al., 2008). Background subtraction methods use a single background every time (static background). However, these methods fail whenever there is a change in light (or scene) conditions. In order to avoid any unfavourable effects due to light change, background frames are updated with incoming frames (dynamic background) (McIvor, 2000). Most background-based moving object detection techniques use gray bands which simplify the algorithm and produce faster results. There also exist techniques which use colour RGB bands (Liu, et al., 2017; Cucchiara, et al., 2003; Nummiaro, et al., 2003; Horprasert, et al., 1999; Carmona, et al., 2008) but their detection methods are sparingly used. Application of RGB bands in moving object detection is less relevant because pixel values in RGB bands are correlated (Piva, et al., 1999).

Some methods also use multiple incoming frames to estimate or model the background. These methods use average filter (Stauffer & Grimson, 1999), median filter (Crnojević, et al., 2009), minimum-maximum filter (Zhou & Jin, 2016), linear predictive filter (Shaikh, et al., 2014), Kalman filter (Subudhi, et al., 2017; Doukas & Maglogiannis, 2008; Chellappa, et al., 2004; Zotkin, et al., 2002), iterative division and correlograms (Han & Bhanu, 2007), single Gaussian (Sadarangani, 2002), mixture of Gaussians (Kamate & Yilmazer, 2015; Piccardi, 2004; Hu, et al., 2004; Han, et al., 2015), and Hidden Markov models (Subudhi, et al., 2017). Neural Network training is also used to update the background frame (Culibrk, et al., 2007). Researchers have also utilized contours from the Canny Edge Detector for moving object detection (Chavez-Garcia & Aycard, 2016; Dedeoglu, 2004; Sezgin & Sankur, 2004).
Background subtraction based moving object detection techniques are computationally expensive and are less suitable for live video streams. Certain background subtraction techniques are used with standalone integrated circuit programmed camera systems (Wolf, et al., 2002). Among all the moving object detection techniques, image thresholding is used to extract a moving object. It plays very critical role as an inappropriate threshold value may fail to remove noise or make the object completely disappear from the frame. Researchers have tried to use different thresholding methods with different moving object detection techniques. Most thresholding techniques are based on image statistics and do not consider spatial distribution of pixels.

1.4.3 No-Reference image quality measurement

An image consists of a single or multiple objects (also known as “foreground”) and backgrounds. Spatial distribution of different objects in an image and their colour distribution plays an important role in image quality perception (Wang, et al., 2002). Full-Reference (FR) and Reduced-Reference (RR) image quality measurement techniques require a reference image which is not available for all the scenes. There are No-Reference (NR) image quality measurement techniques which do not require reference images. In NR image quality measurement, image statistics are used to determine image quality by calculating image RMSE (root mean square error) and PSNR (peak signal to noise ratio) of the image pixels. These parameters are computed from gray images converted from colour images (Wang, et al., 2002; Choi, et al., 2009). It is known that gray images do not necessarily correspond to the visual observations by the human eye. Other techniques used to calculate image quality are fuzzy similarity (Choi & Lee, 2011), vector morphological
operations (Tang, et al., 2011), discrete Fourier transform (DFT), discrete wavelet transform (DWT), and discrete cosine transform (DCT) (Gabarda & Cristóbal, 2007). A new universal image quality index has been proposed (Kang, et al., 2014) which is also computed for gray images. Along with format, the “quality” of an image also depends on brightness, contrast, smoothing, noise, spatial distribution, colours, viewing distance, and viewing angle. Considering these parameters, a human visual system (HVS) model was developed which performed better than RMSE and PSNR quality parameters (Li, et al., 2016). These quantitative quality scores are based on image distortions and do not include human vision model.

1.5 Problem statement

There are various applications of images and videos captured by still and video cameras. Camera outputs are used for moving object detection and are evaluated by human eyes. Existing techniques of moving object detection are either slow or work only under certain scene conditions. Both frame-based and background-based techniques have significant importance with respect to processing speed and object detection rate, but they fail to achieve their purpose of moving object detection. Existing No-Reference image evaluation techniques does not computer image quality score as it is viewed by human eyes and quality scores from different methods are interpreted differently by separate users. The following specific problems were identified with frame-based moving object detection, background-based moving object detection, and quality measurement of camera output according to human eye perception.
a. Being computationally less complex and fast in processing, the frame differencing technique of moving object detection failed to detect the moving object when: 1) the object is moving slowly, and 2) the difference between two consecutive frames is noisy due to various scene conditions. In both situations, it is difficult to determine if the object is detected or not.

b. Background subtraction technique using statistical modeling is computationally expensive and the background needs to be updated with every incoming frame. Most background subtraction techniques use gray images which are obtained from RGB bands. Information is lost from RGB to gray conversion and application of RGB bands is not feasible as they are correlated.

c. Image thresholding is an integral part of moving object detection. Either frame-based or background-based, image thresholding is used with all the techniques. Application of a manual threshold is not feasible because of changing scene conditions and none of the thresholding techniques work under all scene conditions.

d. Image quality quantification is important as human evaluation of camera outputs is not uniform across different users. Both full and reduced reference image quality evaluation techniques require reference images which are not available for all scene conditions. No-Reference image quality measurement techniques are based on gray bands which does not score images as perceived by human eyes.

This PhD dissertation addresses the above mentioned problems and proposes a solution to them in the journal papers mentioned in section 1.1.
1.6 Research Objectives

The objective of this research is to provide effective solutions to the problems mentioned in section 1.5. The major tasks undertaken to accomplish this are as follows:

a. Find an effective solution to the problems involved in detecting moving objects using the frame differencing technique;

b. Develop an independent background subtraction technique which is independent of statistical models and utilize colour bands for the detection of moving objects under different scene conditions;

c. Develop a thresholding methodology using pixel spatial information which works under different scene conditions;

d. Develop a No-Reference image quality evaluation technique which evaluates camera outputs based on human vision. The developed technique should be independent of camera technology and image processing algorithms.

1.7 Data and Metrics

In order to measure robustness of developed algorithms of moving object detection, two different types of cameras are used. These include daily use mobile cameras and high end security camera. Footstep sound recorded from embedded microphones in mobile cameras are used to improve results from frame differencing technique. Synchronized sound recorded with video removed challenges of sound synchronization. For the improvement of background subtraction technique, high end security cameras are used. High end security cameras are different than daily use mobile cameras. Noise distribution
is different in security cameras and these give more flexibility to users to adjust camera gain, saturation and colour balances. For image quality measurement, LIVE, one of the most commonly used image database in image quality measurement is used. Each image in the database comes with a quality score evaluated by different users. Availability of quality score make it easy to compare results from developed technique. Data, instrument, and metrics used to evaluate methods presented in chapters 2 to 4 are summarized in Table 1.1.

<table>
<thead>
<tr>
<th>No</th>
<th>Data</th>
<th>Description</th>
<th>Chapter</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>iPhone 4 Camera (Inc., 2014)</td>
<td>Megapixels: 5MP Video Resolution: 568×320 Video Frame Rate: 30 fps Audio Bit Rate: 65 kbps Audio Sample Rate: 44.1 kHz</td>
<td>Chapter 2</td>
</tr>
<tr>
<td>2.</td>
<td>iPhone 5 Camera (Inc., 2018)</td>
<td>Megapixels: 8MP Video Resolution: 1920×1080 Video Frame Rate: 30 fps Audio Bit Rate: 63 kbps Audio Sample Rate: 44.1 kHz</td>
<td>Chapter 2</td>
</tr>
<tr>
<td>4.</td>
<td>LIVE image database (Sheikh, et al., 2005)</td>
<td>Image dataset from Laboratory of Image &amp; Video Engineering, University of Texas at Austin is available with quality scores from human observations.</td>
<td>Chapter 4</td>
</tr>
</tbody>
</table>
1.8 Overview of each chapter

- **Chapter 1** is an overview of the dissertation presenting the background of the study, selection of research topic, review of existing solutions, problem statement, objective of the study, and an overview of each chapter.

- **Chapters 2 to 4** consists of journal papers which present the major contributions of this research. Specifically:
  
  o **Chapter 2** investigates frame-based moving object detection techniques. Frame-based moving object detection techniques are computationally fast but less often used. This is because these techniques only detect the boundary of moving objects and fail to detect a slow-moving object. Noisy frames further make it difficult to detect moving objects. Moving object (human) detection using frame differencing technique is improved by using footstep sound. Object detection rate is improved when the original technique fails to detect the moving object, partially detects the object, or when the frame is noisy. The developed technique is implemented using videos captured under different scene conditions and the results are post-processed. Footstep sound is detected using Mel-Frequency Cepstrum Coefficients and start-end time stamps of each footstep sound are computed. Detection of footstep sound made it certain that the moving object is present in frames within timestamps. The video frames within the duration of footstep sound are analyzed for a missing/partially detected object and noisy outputs. As an object is better represented by an ellipse, it is modelled as an ellipse which is fitted using a least squares technique on pixels detected as the moving object. For frames with noisy outputs, the object is found by removing noise with spatial segmentation. The
methodology discussed used a camera with in-built microphone and the application of footstep sound depends on it being generated from the moving object.

- **Chapter 3** presents an improved background-based moving object detection technique. Most of the moving object detection techniques based on background subtraction use gray images. Information is lost when an RGB image is converted to gray and application of RGB images is not feasible as the bands are correlated. A new background subtraction technique using Lab colour space is developed. Because thresholding is an integral part of moving object detection and none of the thresholding techniques work under all conditions, a spatial thresholding technique is also developed. Results from the developed technique are better than the results obtained from Otsu and Adaptive thresholding-based background subtraction techniques.

- **Chapter 4** presents a novel No-Reference image quality measurement technique. Existing No-Reference image quality measurement techniques are based on gray images and quality scores generated from these images does not match as the image is perceived by human eyes. Both Full-Reference and Reduced-Reference image quality measurement techniques require reference images which are difficult to find for every camera scene. The developed methodology computes image quality based on spatial distribution of colour segments and their colour distances from the neighborhood segments in the Lab colour space. Results from the developed methodology are compared with other commonly used NR image quality measurement techniques which are not sensitive to colour values. Outputs show that the developed Spatial Colour Referenceless Perceived Image Quality
(SCORPIQ) methodology gives a better image quality score which is in accordance with human eye perception.

- **Chapter 5** summarizes the research and future work.

### 1.9 References


Chapter 2 : Using Mel-Frequency Audio Features from Footstep Sound and Spatial Segmentation Techniques to Improve Frame-based Moving Object Detection

Abstract

Moving object detection in video streams is a challenging and integral part of computer vision which is used in surveillance, traffic and site monitoring, and navigation. Compared to the background-based techniques, the frame differencing technique is computationally inexpensive. However, this technique only detects the boundary of a moving object. Due to changing light conditions, shadows, poor contrast between object and background, and a slow-moving object; the object detection rate from frame differencing technique is less than it would be using other techniques. This is because the number of noisy frames and frames with missing/partially detected object increases. Application of large kernel size morphological operations fails to remove noise as such operations might remove the boundary (or part) of a moving object. In this paper, we propose a methodology to improve the frame differencing technique for a single moving object using footstep sound generated by a moving object. Audio recorded with the video system is processed and footstep sound is detected using audio features computed as Mel-Frequency Cepstral Coefficients. Number of frames within each footstep sound are counted and processed. Spatial segmentation is used to find the moving object in noisy frames. The methodology is demonstrated by post-processing different videos from different scenes where only one

object was moving and producing the footstep sound. An ellipse which fits the moving object better than a rectangle is used to model the object. A missing or partially detected object is recovered by modelling an ellipse using a moving object from other neighbourhood frames.

2.1 Introduction

Moving object detection in video streams is an integral part of computer vision. An object of interest is detected in the video (or image sequence) using a computer program. This topic also has great significance in surveillance, traffic, site monitoring, face detection, robotics, navigation, infrastructure, and military applications (Kumar, et al., 2016; Kamate & Yilmazer, 2015). Moving object detection techniques can be divided into two categories: background-based and frame-based techniques. Background-based techniques are affected by light change, shadows, clutter and camouflage. Statistical modeling is used to compensate for the changing light conditions which increased the algorithm complexity and demanded for additional computational power for efficient processing. On the other hand, since frame-based techniques use consecutive frames, these does not require background modeling and hence are computationally inexpensive. Frame-based techniques are less complex, better in susceptibility to light change and less affected by shadows (Roshan & Zhang, 2014). “Frame differencing” and “Optical flow” are the two most commonly used frame-based techniques of moving object detection where frame differencing technique is less expensive than Optical flow technique of moving object detection. Absolute value of difference between two consecutive frames is subjected to threshold in frame differencing technique whereas Optical flow velocity is computed
between two consecutive frames is subjected to threshold in Optical flow moving object detection technique.

Though computationally less complex and fast in processing, the frame differencing technique has not been used much mainly because of poor outputs. This technique detects only the boundary of the moving object. For slow moving objects, this technique either fails to detect or only partially detects the moving object, which makes it hard to determine if the entire object is detected or not. The relative success of this technique is also influenced by poor contrast between object and its background (stationary or dynamic), camouflage, clutter, noise, and shadows. Multiple light sources are used in indoor conditions causing multiple shadows and reflections. Under all these conditions, frame differencing technique produces a noisy object, where the noise is distributed all over the frame. This makes it quite difficult to find the moving object. This paper is focused on improvements of the frame differencing technique for detection of a moving object (e.g., a human). Human detection is referred to as “moving object detection” in the rest of this article.

In certain conditions, a moving object (a human) produces footstep sound as it moves. Most modern video recording camera systems come with an inbuilt microphone. Some cameras come with omnidirectional and some with stereo (or bidirectional) microphone which record sound present in the surroundings. Presence of footstep sound indicates the presence of an object. In a novel technique developed in this paper, the footstep sound is used to improve the frame differencing technique whenever it fails to detect a moving object due to various reasons. The camera’s inbuilt microphone system records sound which is synchronized with the video. The recorded sound is processed and footsteps are
detected. Presence of footstep sound is used as a confirmation of the presence of moving object in video frames even if moving object detection technique has failed to detect the object. After the confirmation of presence of object using footstep sound, frames adjacent to the frame of interest are used to model the moving object (Figure 2.1).

**Figure 2.1.** Process showing the improvement of Frame Differencing (FD) moving object detection technique using footstep sound recorded with video recording system

Audio amplitude and background noise varies from one camera scene to another. For successful detection of footstep sound, high frequency background noise is filtered out
using low pass filter. Audio features from processed audio data are computed as Mel Frequency Cepstral Coefficients (MFCC) (Logan, 2000). The coefficients include audio features of all frequencies. Lower order MFCC coefficients corresponding to low frequency audio features are used and processed to detect footstep sound. The number of video frames within the detected footstep sound are counted and analyzed for the presence of the moving object.

The proposed methodology discussed in this paper improves the moving object detection rate of frame differencing technique by post processing captured videos under different scene conditions. A frame where object is partially detected or completely vanished due to noise and other scene conditions, spatial segmentation is used to detect the moving object in that frame. This is achieved by taking into consideration of neighbouring frames where an object is detected. A partially detected object refers to any extent of incomplete detection of the moving object. As a result, application of footstep sound results in identifying a greater number of frames with the detected moving object. This can also be extended for other frame and background-based moving object techniques.

2.2 Literature Review

A significant amount of research has been conducted to date focusing on the detection of moving objects using videos from different cameras. However, very limited data is available on the applications of footstep sound to improve the shortcomings of frame differencing method. Background subtraction (McIvor, 2000) and frame differencing (Cui, et al., 2012) are two basic techniques used for moving object detection. Background subtraction uses a video frame without an object as background and subtracts each
incoming frame from the background. The frame differencing technique computes the
difference between two consecutive frames and finds the moving object. Due to the
presence of noise, slow moving objects, shadows, changing background, poor contrast
between object and background, dynamic background, clutter, and camouflage, (Shaikh,
et al., 2014) final outcomes were not useful. These basic techniques have been further
developed. Optical flow (Shafie, et al., 2009), Gaussian mixture model (Stauffer &
Grimson, 1999), and wavelet transform based method (Crnojević, et al., 2009) are some of
the other most commonly used moving object detection techniques. There might be various
reasons to choose one technique over another but, due to different variables and conditions,
it is hard to find a single technique suitable for all conditions. Researchers are still working
on improving these techniques. Recently, Zhou and Jin (2016) used two-dimensional
principle component analysis to update the background frame (Zhou & Jin, 2016) and
Subudhi et al. (2017) used spatio-temporal multilayer compound Markov Random Field
(Subudhi, et al., 2017) for histogram based moving object detection.

A video-only based moving object detection technique uses different properties from
video frames to detect the moving object. Some use two or three consecutive frames to
detect the change and others use static background with incoming frames. Detailed
literature on most of the basic moving object detection techniques is available (McIvor,
2000; Shaikh, et al., 2014; Piccardi, 2004; Hu, et al., 2004). Gaussian Mixture Model
(GMM) is one of the most popular background subtraction techniques because of its
susceptibility to adopt light change and shadows (Piccardi, 2004). This has been further
improved by using colour information (Sadarangani, 2002; Han, et al., 2015). An
automatic graph cut method has been introduced to detect a moving object which uses
frame difference of (or between) consecutive frames as the seed (or initial detection) (Cui, et al., 2012). This method will fail when there is no seed due to a slow-moving object or when the seed region is large due to noise in the frame. A neural network is implemented to learn about background change and track the moving object (Song, et al., 2015; Qin, et al., 2016). This requires neural network training when the scenario is altered. Researchers have also fused two or more techniques to obtain improved final outputs (Siebel & Maybank, 2002; Zhou & Zhang, 2005; Heikkila & Pietikainen, 2006).

Besides using videos, researchers have also considered sound as a variable for moving object detection. Although sound-only techniques of moving object detection have not been used in computer vision, they are used to detect a falling person and provide medical help to the elderly (Doukas & Maglogiannis, 2008). A sound array is used to detect and track a moving object in traffic, and to monitor pedestrians. Drawbacks of sound-only methods are that the precise location of a moving object is unknown and tracking of an object requires a larger network of audio sensors. Sound sensors are also used with video in multimodal systems to track the speaking person in the camera scene such that the camera is focused on the person (Chellappa, et al., 2004). Multi-modal systems are further developed by using multiple video and audio sensors to detect and track the moving object in the scene (Zotkin, et al., 2002). Other sensors such as infrared video, radar and LiDAR have also been used with optical imaging systems in multi-modal systems (Han & Bhanu, 2007; Chavez-Garcia & Aycard, 2016). With the increasing number of sensors in multimodal systems, there is also an increase in the complexity of these systems. Neither the multimodal systems nor video-only based techniques have addressed the problem of missing or partially detected moving objects.
2.3 Methodology

First, the frame differencing technique of moving object detection is used to detect a moving object in video frames. Later, sound recorded with the video recording system is processed and footstep sound is detected. Duration, start and end time of each detected footstep is calculated and used to improve the moving object detection rate using a frame differencing technique. Spatial segmentation is developed to find the moving object in noisy frames and missing or partially detected object is modelled as a complete object using neighbourhood frames to the frame under consideration.

2.3.1 Moving Object Detection

A “frame differencing” technique subtracts two consecutive frames and applies a threshold to the absolute difference (Eqn 2.1).

\[
\text{Output Pixel Value} = \begin{cases} 
1 & \text{if } |f_t - f_{t-1}| > Th \\
0 & \text{Otherwise}
\end{cases}
\]  

(2.1)

Where \(f_t\) is frame at time \(t\), \(f_{t-1}\) is frame at time \(t-1\), and \(Th\) is threshold. The threshold value plays a critical role (Dedeoglu, 2004) and should be updated with change in the camera scene. Different techniques exist for image thresholding for multi-band (or colour) and gray images (Sezgin & Sankur, 2004). The gray image histogram-based Otsu method of automatic threshold (Otsu, 1979) and manual threshold are popular in moving object detection. Manual selection of threshold is critical as it may differ for different objects and camera scenes. Only one manual threshold is used for the entire dataset whereas each absolute difference of consecutive frames uses different threshold when the automatic Otsu
method of thresholding is applied. Both the manual and Otsu thresholding methods as applied to dataset and results are discussed in section 2.4.

2.3.1.1 Morphology: Random noise is present in original video frames and its spatial location changes from frame to frame. For a given threshold value, depending on scene condition, the amplitude of noise also changes from one frame to another. These unwanted noise pixels are removed (or minimized) using morphological operations. The most commonly used morphological operation to reduce noise in moving object detection is erosion followed by dilation (Dedeoglu, 2004). For a square kernel of 3×3 is used in morphological operations, erosion removes one pixel from the boundary of an object and dilation ads one pixel along the boundary. This sequential operation removes the noisy pixels by making it sure that partial or complete boundaries of detected object are least affected.

2.3.1.2 Ellipse Fitting: Most moving object detection techniques represent a moving object by its bounding rectangle. However, a moving object is better represented as an ellipse than a rectangle. A least-squares ellipse is fitted to the group of pixels detected as a moving object. In order to find a bounding ellipse, object boundary pixels are found as convex hull points. MATLAB command ‘convexhull’ (Inc. 2019), is used to find object boundary pixels and least squares ellipse $E(x, y, a, b, \theta)$ is modelled (Gal, 2003). On the ellipse $E$, $x$: x center, $y$: y center, $a$: semi-major axis, $b$: semi-minor axis and $\theta$: orientation angle. When the pixels detected as a moving object are randomly distributed all over the frame, the frame is considered noisy and a noise removal algorithm (section 2.3.4) is applied.
2.3.2 Audio Processing for Footstep Sound Detection

Sound recorded with video is extracted and processed to detect footsteps (Figure 2.2). Noise and unwanted sounds present in the surroundings are also recorded. In order to minimize high frequency noise produced by the surrounding environment, a low pass filter is used. Audio features of filtered audio wave are extracted as MFCC coefficients. These MFCC coefficients are used to detect footstep sound, along with their locations (or times) and durations with start and end timestamps. Significance of the first MFCC coefficient in Figure 2.2 is explained in the following sections.

![Diagram of audio processing and video frame counting for each peak location in input sound](image)

**Figure 2.2.** Audio processing and video frame counting for each peak location in input sound

2.3.2.1 Low Pass Filter: High frequency background noise recorded with camera microphone system is reduced by applying low pass filter to the audio data. A low pass filter with rational transfer function (Oppenheim, et al., 1999) is applied using the command ‘filter’ in MATLAB. Filter uses 0.05 as numerator coefficient which is equivalent to averaging of audio signal with a window size of 20 samples.

2.3.2.2 MFCC Coefficients: Filtered audio data is used and audio features are extracted as MFCC coefficients which are computed using MATLAB (Wojcicki, 2011). Audio features
are computed for each 200ms audio frame (audio is divided into audio frames). 13 (0 – 12) MFCC coefficients represent audio features are computed for each frame. These coefficients are arranged in increasing order of frequencies where the zeroth coefficient is the average of 12 coefficients. Lower-order coefficients (e.g. 1st MFCC coefficient) represent low-frequency audio features and higher-order coefficients (e.g. 12th MFCC coefficient) represent high-frequency audio features (Logan, 2000).

2.3.2.3 1st MFCC Coefficients: A normal moving human produces a footstep sound each 0.25 – 1 second. In terms of frequency, footsteps are generated at a frequency of 1 – 4 Hz. This is low-frequency sound and lower order MFCC coefficients are useful. 1st MFCC coefficients of audio features are used to detect footstep sound. Time vs 1st MFCC coefficients plot shows that lower peaks (i.e., those below the average 1st coefficients) correspond to the time when a footstep sound is produced (Figure 2.5). These peaks are detected and their start and end times are calculated as described in the next section.

2.3.2.4 Peak Locations: The average of First MFCC coefficients is calculated and the coefficients values below average are found. These coefficients form peaks of lower half of a sinusoidal wave starting from the average line of coefficients. Start and end timestamps for each peak are computed at the average line. Coefficients are discrete in time, and may not have a value at average line. In these cases, start and/or end timestamps of peaks are estimated by finding the intersection point of two lines: one as the average line and second the line joining the two adjacent coefficients above and below the average. The duration of footstep sound is computed using start and end timestamps of detected peaks.
2.3.3 Video Frames Counting

A video is captured at 30 frames per second and audio is sampled at 44.1 kHz. The audio sampling rate is 1470 times higher than the video frame rate. The timestamp of each frame in the video is calculated using the video frame rate. Start and end timestamps for each footstep are known and these are later fused with the video frames timestamp. This gives the number of video frames which fall within the start and end timestamps of each footstep sound (Figure 2.2).

2.3.4 Missing, Partially and Noisy Detected Object

Due to the slow movement of an object, light change, random noise, light source, threshold, and other camera scene conditions, the output object can be partially detected, not detected at all, noisy, or a combination thereof. Under these conditions, it is difficult to confirm the presence of a moving object in certain frames. However, the presence of footstep sound helps to detect a moving object. This section describes the methodology used to improve the detection of a moving object in frames with a noisy, missing/partially detected object.

2.3.4.1 Missing/Partially Detected Moving Object: It is natural that an object does not move at constant speed over time and the object movement itself is non-uniform. The slow and non-uniform movement of an object and/or non-ideal value of threshold used in image thresholding results in a missing/partially detected object. When presence of a footstep sound is confirmed and a frame \(f_t\) with a missing/partially detected object is found within the time duration of footstep sound, an object in frame \(f_t\) is modelled with the help of
neighbouring frames $f_{t-1}$ and $f_{t+1}$ with a fully detected object. Ellipses $E_{t-1}(x, y, a, b, \theta)$ and $E_{t+1}(x, y, a, b, \theta)$ are fitted to the points detected as moving object in frames $f_{t-1}$ and $f_{t+1}$ respectively. Next, the object in $f_t$ is modelled as an average ellipse $E_t(x, y, a, b, \theta)$ of $E_{t-1}$ and $E_{t+1}$. If the frame $f_{t+1}$ does not contain a fully detected object or does not fall within the duration of present footstep sound, the next frame with a complete detected object is considered in modelling the moving object as an ellipse $E_t$ in frame $f_t$. If the next frame with a complete moving object is not found in frames detected within the footstep sound, object modelling is terminated and moved to the frames within the next footstep sound. For real-time applications, actual detection of moving objects might lag by a certain number of frames (or time) as ellipse modelling in the frame where the object is either missing or partially detected requires a frame with a complete moving object.

2.3.4.2 Noisy Output: Different reflections and multiple shadows due to multiple light sources close to the object cause rapid light changes in the camera scene. These generate noisy outputs which cannot be completely removed using the morphological operations applied during moving object detection (section 2.3.1). Inappropriate threshold values can also generate noisy outputs. A spatial segmentation technique is developed as part of this research to remove this noise.

Morphological Operations: A large kernel size (5×5 or 7×7 or higher) structuring element is required to remove the accumulated noisy pixels. A frame-based moving object detection technique that detects only the object boundary and application of a large kernel size will make the detected object disappear (or partially disappear). Due to this limitation,
a 3x3 kernel is used for morphological operations and “salt and pepper” image noise is further reduced.

*Spatial Segmentation:* Due to the limitations of morphological operations, noise is not completely removed from the frame of a detected object. It is assumed that noise is distributed across the frame and multiple pixels belonging to the moving object are close to each other. Pixels in noisy frame are spatially segmented using K-means clustering (Arthur & Vassilvitskii, 2007).

*Clustering and Weighting:* K-means clustering is applied to the detected pixels and the image frame is divided into 20 clusters. The number of pixels in each cluster and their mean location (row, column) are determined. A weight (Eqn 2.2) is assigned to each cluster based on the number of pixels it contains. Using the assumptions made earlier under Spatial Segmentation in Section 2.3.4.2, three clusters with maximum weight are found. All the pixels contained in these three clusters are combined and their \( x \) and \( y \) standard deviations (\( s_x \) and \( s_y \)) are computed.

\[
W_{\text{cluster}} = \frac{\text{Number of detected pixels in the cluster}}{\text{Total number of detected pixels in frame}}
\]  

(2.2)

*Noise Removal:* A cluster belonging to boundary pixels in the object has more weight than other clusters with noise pixels only. If the whole object is not within one cluster, part of the object is contained in the neighbouring clusters. The Maximum Weight among all clusters is found (\( W_{\text{max}} \)) and two-dimensional standard deviation (\( s_x \) and \( s_y \)) calculations are used to find the clusters which belong to the moving object (Eqn 2.3).
Moving object cluster = \begin{cases} 
Yes & \text{if } \left\{ \begin{array}{l}
\text{Cluster weight} > 0.05 \times W_{\text{max}} \\
\text{and}
\text{Cluster distance} < 3 \times D_{\text{th}}
\end{array} \right. \\
No & \text{Otherwise}
\end{cases} 
\tag{2.3}

Where the threshold distance, \( D_{\text{th}} = \sqrt{s_x^2 + s_y^2} \).

Threshold weight is 5% of maximum weight and threshold distance is 3 times the two-dimensional standard deviation of three clusters with maximum weight.

2.4 Results and Discussion

2.4.1 Camera System

Videos of a moving object were captured in indoor conditions using an iPhone 4 and 5 mobile cameras (Table 2.1). The iPhone’s inbuilt microphone system was used to capture audio generated from the moving object.

<table>
<thead>
<tr>
<th>Camera</th>
<th>iPhone4, 5MP still camera (Inc., 2014)</th>
<th>iPhone5s, 8MP still camera (Inc., 2018)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Video Resolution</td>
<td>568 \times 320</td>
<td>1920 \times 1080</td>
</tr>
<tr>
<td>Video Frame Rate</td>
<td>30 fps</td>
<td>30 fps</td>
</tr>
<tr>
<td>Audio Bit Rate</td>
<td>65 kbps</td>
<td>63 kbps</td>
</tr>
<tr>
<td>Audio Sample Rate</td>
<td>44.1 kHz</td>
<td>44.1 kHz</td>
</tr>
</tbody>
</table>
2.4.2 Camera Scene

Four indoor scenes were selected and video with a single moving object was captured. Regular fluorescent light tubes were used as a light source. The object in each camera scene started moving from the far end and began moving towards the camera in a zig-zag pattern with normal speed (Figure 2.3). Video of a moving object in outdoor conditions can also be used if that moving object’s footstep sound is detectable distinct from the background noise (traffic, wind, music etc.).

![Image of camera scenes]

**Figure 2.3.** Camera scenes in captured videos  
(a) Video 1: Object moving in indoor room 1  
(b) Video 2: Object moving in indoor room 2  
(c) Video 3: Object moving in indoor hallway 1  
(d) Video 4: Object moving in indoor hallway 2

2.4.3 Manual and Otsu Thresholding for Frame Differencing

A frame differencing technique of moving object detection is applied to the captured videos. As discussed in section 2.1, threshold plays an important role in moving object detection. Both manual (using trial and error) and automatic (using Otsu method) thresholds are applied (Figure 2.4) and moving object detection results are evaluated. Different threshold values for each absolute difference of consecutive frames are used in Otsu method (determined automatically using Otsu algorithm (Otsu, 1979)) whereas a single threshold value is used for all absolute difference of consecutive frames in manual thresholding. The optimal value of manual threshold is selected such that the object is detected successfully with minimal noise. The Otsu method requires less user interaction and determines threshold values based on frame difference, but the results show that object
detection rate is better using manual thresholding than Otsu thresholding. Sample frames from four videos are shown in Figure 2.4. As compared to manual thresholding, a large number of output frames are noisy and object is partially detected using Otsu threshold. As a result, manual thresholding is used with frame differencing technique and final outputs are analyzed based on it.

Figure 2.4. Frame differencing moving object detection results using manual and Otsu thresholding. **Row 1:** manual threshold and **Row 2:** Otsu threshold (Frames from left to right: Video 1, frame 30; Video 1, frame 184; Video 2, frame 30 and Video 2, frame 254). **Row 3:** manual threshold and **Row 4:** Otsu threshold (Frames from left to right: Video 3, frame 147; Video 3, frame 482; Video 4, frame 132 and Video 4, frame 214)
2.4.4 Footstep Sound and Frame Counting

An audio signal (Figure 2.5) is extracted from each of the four videos and processed for footstep sound detection. Noise from each audio signal is removed by applying a low-pass filter and the respective MFCC coefficients of each audio feature are computed.  

![Table 2.2 Frame counts in videos during footstep sound and frame differencing detection analysis](image)

<table>
<thead>
<tr>
<th>Description</th>
<th>Video 1</th>
<th>Video 2</th>
<th>Video 3</th>
<th>Video 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Video duration (sec)</td>
<td>11.07</td>
<td>20.10</td>
<td>19.12</td>
<td>13.62</td>
</tr>
<tr>
<td>Number of footsteps</td>
<td>20</td>
<td>32</td>
<td>35</td>
<td>25</td>
</tr>
<tr>
<td>Total number of video frames</td>
<td>354</td>
<td>605</td>
<td>576</td>
<td>441</td>
</tr>
<tr>
<td>Number of frames within footstep</td>
<td>166</td>
<td>260</td>
<td>205</td>
<td>167</td>
</tr>
<tr>
<td>Number of frames with complete object within footstep</td>
<td>60</td>
<td>176</td>
<td>91</td>
<td>77</td>
</tr>
<tr>
<td>Failure rate (%) of frame differencing within footstep</td>
<td>64%</td>
<td>32%</td>
<td>56%</td>
<td>54%</td>
</tr>
</tbody>
</table>

![Figure 2.5. Video 1 footstep sound waveform and detection of peak locations corresponding to footstep sound using first MFCC coefficients](image)
MFCC coefficient values are processed using the methodology discussed in sections 2.3.2 and 2.3.3. For each audio signal, start and end timestamps of each footstep sound are fused with the timestamps of video frames detected using frame differencing technique. Thereafter the number of frames during each footstep sound are counted (Table 2.2). Based on this analysis, 166 out of 354, 260 out of 605, 205 out of 576 and 167 out of 441 frames fall within the duration of detected footstep sounds for Videos 1 to 4 respectively. This confirms the presence of moving object in these frames. Further; only 60, 176, 91 and 77 frames out of 166, 260, 205 and 167 frames respectively are detected with a complete moving object. It shows that 64%, 32%, 56% and 54% frames within the duration of footstep sounds for Videos 1 to 4 respectively are either noisy or with a missing/partially detected object or a combination thereof (Figure 2.6). These frames are further processed and the missing or partially detected object is modelled.

**Figure 2.6.** Frame differencing output frame analysis based on fully detected object falling within the duration of detected footstep sound
2.4.5 Modelling Moving Object in Frames with Partially Detected, Missing Object and Noisy Output

There are frames with a missing object, a partially detected object, and noisy output. An object is modelled to the frames which have a missing or partially detected moving object and which fall during the presence of footstep sound. Figure 2.7 shows three consecutive frames each from Videos 1 to 4. All three frames occurred during the presence of footstep sound and the frame differencing technique partially detected the moving object in Frame 167 of Video 1, Frame 349 of Video 2, Frame 360 of Video 3 and Frame 172 of Video 4. In most of the cases, the object in these frames is ignored as noise or eliminated from morphological operations. As discussed in section 2.3.4, a partially detected object in Frame 167 of Video 1 is modelled as a complete object (Figure 2.7) using the object detected in neighbouring Frames 166 and 168 (section 2.3.4). When one (or both) of the neighbouring frames to the frame under consideration (e.g. Frame 359 and 361 of Video 3 in Figure 2.7) is noisy, the noise is first removed using spatial segmentation. If the subsequent frame fails to detect an ellipse or does not contain a complete object, the next available frame with a detected object is used. For real time application, moving object detection will be delayed until next frame is detected with a complete moving object.

A noisy output frame generated from the frame differencing technique makes it difficult to find a moving object present in the frame. Noise is removed using spatial segmentation discussed in section 2.3.4. Figure 2.8 shows the noisy frames 187, 3, 570 and 314 from videos 1 to 4 respectively. These frames occurred during the detected footstep sound. If noise is not removed, object is detected as an entire image frame containing all
the noisy pixels. After applying morphological operations and spatial segmentation, the noise is removed and an ellipse is fitted to represent the moving object.

**Figure 2.7.** Moving object detection and modelling of missing/partially/noisy moving object using neighbouring frames. Three consecutive frames within the presence of footstep sound are shown. Object in the middle frame is modelled using previous and next frames. **Rows 1 & 2:** Video 1, Frames 166, 167 and 168. **Rows 3 & 4:** Video 2, Frames 348, 349 and 350. **Rows 5 & 6:** Video 3, Frames 359, 360 and 361. **Rows 7 & 8:** Video 4, Frames 171, 172 and 173
Using frame differencing technique, 64%, 32%, 56% and 54% of frames (within the presence of footstep sound) of Videos 1 to 4 respectively are found to be noisy or with missing/partially detected outputs or a combination thereof. Footstep sound is used to confirm the presence of a moving object in these frames. The moving object is then modelled as discussed in section 2.3.4. Figure 2.9 shows video frames with moving object modelled as an ellipse. All the frames are extracted during the presence of footstep sound in respective videos. Neighbourhood frames with a complete detected object are used to model the object (as an ellipse) in frames with the missing/partially detected object. The object is extracted and modelled as an ellipse for noisy frames using spatial segmentation.

**Figure 2.8.** Noisy frames and noise removal using spatial segmentation. **Row 1:** Frame 187 of Video 1, **Row 2:** Frame 3 of Video 2, **Row 3:** Frame 570 of Video 3 and **Row 4:** Frame 314 of Video 4. **Column 1:** Frame differencing output frame, **Column 2:** Spatially segmented pixels, **Column 3:** Frames after removing noise, and **Column 4:** Object modelled after noise removal

**2.4.6 Moving Object Detection**
As discussed in section 2.2, similar problems of missing/partially detected moving object and noisy outputs also exist with other frame-based methods of moving object detection. Object detected video frames with a missing, partial, and noisy output frames using optical flow and wavelet transform based moving object detection techniques are shown in Figure 2.10.


A missing, partial, and noisy output from optical flow and wavelet transform based techniques can also be improved by incorporating footstep sound into the detection process. While different moving object detection techniques have different complexities and take different processing times; processing steps and computation time for the application of footstep sound remains the same. For Video 1, out of the 166 frames occurring during the presence of footstep sound, a complete object is detected in 105 frames using optical flow-
Based technique, in 117 frames using wavelet transform based technique, and in 60 frames using frame differencing technique. With the application of footstep sound, a complete moving object is detected in all 166 frames of Video 1. Similar analysis is applied to Videos 2, 3 and 4 which leads to the conclusion that – on average – the frame differencing technique is improved by 52% in comparison to 31% of optical flow and 45% of wavelet transform based techniques (Table 2.3).

**Table 2.3** Improvements to the frame-based moving object detection techniques when footstep sound is detected

<table>
<thead>
<tr>
<th>Description</th>
<th>Frame Differencing</th>
<th>Optical Flow</th>
<th>Wavelet Transform</th>
</tr>
</thead>
<tbody>
<tr>
<td>Video</td>
<td>1 2 3 4</td>
<td>1 2 3 4</td>
<td>1 2 3 4</td>
</tr>
<tr>
<td>Object detected frames within footstep sound</td>
<td>60 176 91 77</td>
<td>105 202 116 129</td>
<td>117 142 83 92</td>
</tr>
<tr>
<td>Video frames within footstep sound</td>
<td>166 260 205 167</td>
<td>166 260 205 167</td>
<td>166 260 205 167</td>
</tr>
<tr>
<td>Improvement (%)</td>
<td>64% 32% 56% 54%</td>
<td>37% 22% 43% 23%</td>
<td>30% 45% 60% 45%</td>
</tr>
<tr>
<td>Average Improvement (%)</td>
<td>52%</td>
<td>31%</td>
<td>45%</td>
</tr>
</tbody>
</table>

**Figure 2.10.** Output frames of moving object detection using optical flow and wavelet transform based moving object detection techniques. Sample video frames within duration of footstep sound from Videos 1 to 4. **Row 1:** video output frames using optical flow and **Row 2:** video output frames using wavelet transform.
As a result, frame differencing technique performs poorly in moving object detection when compared to other frame-based techniques. This observation significantly improved with the incorporation of footstep sound. Detection of footstep sound makes it certain that the object is present in the camera scene and the results are based only on the duration of detected footstep sound. Remaining frames whose timestamps do not coincide with the footstep sound are not considered. The performance and results may vary among different video and scene conditions.

2.5 Conclusions

Frame differencing technique based moving object detection for a single object is improved by using footstep sound when the former fails to detect the moving object, partially detects the object, or when the frame is noisy. By computing start and end timestamps of each footstep sound, it is made certain that the moving object is present irrespective of frame differencing results. The video frames within the duration of footstep sound are found and analysed for a missing/partially detected object and noisy outputs. A missing/partially detected object in frame \( f_t \) is modelled as a complete object using the ellipses from neighbourhood frames \( f_{t-1} \) and \( f_{t+1} \). An object is modelled as an ellipse which is fitted using the least squares technique on pixels detected as moving object. When object is not detected in neighbouring frames, the next available frame with the fitted ellipse (complete detected object) is used. For frames with noisy outputs, the object is found by removing noise with spatial segmentation. Application of spatial segmentation to remove noise and model the missing object can also be extended for frames between two
consecutives footsteps if the presence of a moving object is confirmed. As a result, frame differencing technique can be further improved.

The methodology discussed in this paper is demonstrated by detecting moving objects for offline processing and the videos are captured using cameras with inbuilt microphones. The application of footstep sound also depends on it being generated from the moving object. There are circumstances when footstep sound is generated from “off-scene” objects. The accuracy and certainty of the presence of a moving object in a camera scene can be further improved by using advanced directional microphones, which can record sound only from the camera scene. This may require additional footstep sound analysis, sensors, or computation power to confirm the presence of a moving object in the camera scene. The method discussed in this paper includes some of the steps which might be expensive and their applicability for real-time applications need further investigation. The developed technique can also be used with other moving object detection techniques (including background-based) depending on the application, algorithm complexity, and payload. This technique can be extended for multiple moving objects in the camera scene which produce footstep sound at a high frequency. Higher frequency footstep sound can be analyzed using higher order MFCC coefficients.

2.6 Acknowledgements

This research was sponsored by the Atlantic Innovation Fund (AIF) of the Atlantic Canada Opportunities Agency (ACOA) with Yun Zhang as the Principal Investigator.
2.7 References


Chapter 3: Lab Colour Space and Spatial Correlation based Moving Object Detection

Abstract

Background subtraction techniques of moving object detection are commonly used in computer vision programs. Most types of the background subtraction techniques use gray images in which the colour information is lost when converted from RGB images. Application of RGB bands in moving object detection is not feasible as RGB bands are correlated. Apart from the use of gray images which simplify algorithm complexity, researchers have also used other kinds of colour spaces. In a camera scene, outputs change rapidly with changing scene conditions. Consequently, the image background needs to be updated to compensate for light change. To achieve this, certain statistical models are used to model the background which is used in some background subtraction techniques. However, as a combined effect of varying scene conditions and inappropriate image threshold value selection, most of the background subtraction techniques does not give desired results. Multiple image thresholding methods generate different outputs and it is very challenging to find a method working under different scene conditions. In this study, authors have utilized Lab colour space with background subtraction technique and developed image thresholding methods using spatial auto-correlation. Effect of light change is eliminated by using red-green and blue-yellow opposite colour pairs of Lab colour space. As a result, if an object is present in the scene, it has a higher chance of being detected at the same location in the difference of opposite colour pairs. In image

---

thresholding using spatial auto-correlation, pixels with high spatial auto-correlation are extracted as moving objects. With different camera scenes, light conditions, and objects, the basic moving object detection technique is improved which gives better results (by 77%) than gray image-based background subtraction using existing thresholding methods.

3.1 Introduction

Moving object detection techniques are used to detect moving objects present in the camera scene. The detected object can be used in traffic monitoring, security surveillance, site monitoring, face detection, military applications, photography, and robotics (Kamate & Yilmazer, 2015; Kumar, et al., 2016). These techniques are divided into two major categories: Background-based and frame-based techniques. Background-based techniques use a static background and detect the moving object by subtracting it from the incoming frame (Eqn. 3.1). Among all the background-based techniques, background subtraction is the most commonly used. Frame-based techniques use two or more consecutive video frames and are less popular due to their poor outputs. Changing scene conditions due to change in the camera scene illumination, noise, multiple light sources, colour distribution of foreground and background objects, object shadows and the type of light source used to illuminate the camera scene are some major challenges in moving object detection (Yazdi & Bouwmans, 2018). To solve these challenges, different techniques of background subtraction have been developed. These techniques used gray, $RGB$ (R: red, G: green and B: blue), $HSV$ (H: hue, S: saturation and V: value), $YUV$ (Y: luma, UV: chrominance) and $Lab$ (L: lightness, a: red-green opposite colour pair, and b: blue-yellow opposite colour pair) colour spaces. Among all the background subtraction techniques, background modeling using statistical models is very popular. Statistical models are used to update the
background and adopt scene changes due to scene illumination variation (Goyal & Singhai, 2018). A detailed literature on moving object detection techniques is available (McIvor, 2000; Shailh, et al., 2014; Piccardi, 2004; Hu, et al., 2004).

\[
Moving Object = \begin{cases} 
1, & |Background - Foreground| \geq \text{Threshold} \\
0, & \text{Otherwise} 
\end{cases} \quad (3.1)
\]

Background modeling was introduced by Wren et al. by modelling each pixel as a Gaussian distribution (Wren, et al., 1997). Stauffer and Grimson modelled each pixel using a mixture of Gaussians which is known as a Gaussian Mixture Model (GMM) (Stauffer & Grimson, 2000). Background modeling evolved when researchers used nonparametric structures (Elgammal, et al., 2002), pixel’s spatial values (Sheikh & Shah, 2005), block-based structures (Ke, et al., 2011; Savas, et al., 2018), hierarchical block-based structures (Yeh, et al., 2014), and block-based Robust Principal Component Analysis (RPCA) (Yang & Zou, 2015) coupled with GMM. As a result, complexity of background modeling algorithms increased which resulted in longer processing times and the requirement of faster processing systems. Recently, Savas et al. used Kernel Density Estimation model (Elgammal, et al., 2002) to improve background models using pixel blocks (Savas, et al., 2018). Goyal and Singhai modelled the background using local binary patterns for texture-based self-adaptive moving object detection (Goyal & Singhai, 2018).

In order to reduce the complexity of algorithms, most background subtraction techniques use gray images. However, gray images which are obtained by converting \textit{RGB} images lose colour information on conversion. Researchers have also used \textit{RGB} bands for moving object detection. Horprasert et al. developed a computational colour model by measuring brightness and chromaticity distortions using \textit{RGB} images (Horprasert, et al., 1999). Using image pixel values in \textit{RGB} colour space, Cucchiara et al. combined statistical
assumptions with the object-level knowledge of moving objects (Cucchiara, et al., 2003) and Carmona et al. transformed the pixel values to a 2D system representing angle and module of the pixel (Carmona, et al., 2008). Both techniques focused on removing shadows and unwanted noise detected with moving objects in $RGB$ colour space. A background subtraction technique using $RGB$ colour model is presented by Chun-Yang et al. and Bin et al., who improved background model using codebook model (Chun-yang, et al., 2013; Bin, et al., 2017). However, $RGB$ bands of an image are correlated (Piva, et al., 1999) which can be a potential reason behind many studies showing less improvement using $RGB$ bands.

In recent studies, Conventional Neural Networks (CNN) and Deep Learning methods are also utilized to detect moving objects (Luo, et al., 2018; Li, et al., 2018). Both CNN and Deep Learning algorithms have the capability to learn from detection results, but these methods require extensive training and processing powers which increase the associated cost.

Onboard processing using Field Programmable Gate Array (FPGA) (Lopez-Bravo, et al., 2013) is also used in moving object detection which brings new challenges of camera compatibility, data transfer, and equipment cost. Researchers have also utilized synchronized RGBD (Red, Green, Blue plus Depth) sensors to detect moving objects (Maddalena & Petrosino, 2018). RGBD sensors provide depth information in a camera scene.

Removal of shadows, noise, and ghost pixels is very challenging in moving object detection. Different colour spaces have been used to achieve this task. Zhiyong et al. removed object shadows using $Lab$ colour space (Zhiyong, et al., 2017), Hernández-López et al. used a Kinect depth sensor with $Lab$ colour space (Hernández-López, et al., 2012),
and Balcilar et al. detected moving objects using Lab colour space with spatial and temporal smoothing (Balcilar, et al., 2014). HSV colour space is also used in the detection of moving objects and removal of shadows (Cucchiara, et al., 2001). Farou et al. used RGB, HSV and YUV color spaces to remove shadows detected with moving objects obtained from GMM (Farou, et al., 2017). Availability of images and videos in Lab color space also depends on the video being captured in colour mode as majority of security cameras switch to monochrome mode under low light conditions (Kruegle, 2007). However, with the use of advanced camera technology and image fusion techniques, colour videos can be obtained under very low light situations (Yun, et al., 2016). Hence, colour pairs are available for video streams and moving object detection. Lab colour space decomposes an RGB image into “lightness”, “red-green”, and “blue-yellow” opposite colour pairs (ITU-R, 2017) and the effect of light change in moving object detection can be eliminated by detecting moving objects using opposite colour pairs of Lab colour space.

In every moving object detection technique, image thresholding is used to extract moving objects and remove unwanted noise (Eqn. 3.1). Different image thresholding methods have been developed and used by researchers (Sezgin & Sankur, 2004; Carabias, 2012; Otsu, 1979; Kapur, et al., 1985; Kittler & Illingworth, 1986; Bradely & Roth, 2007), but it is difficult to find one single method working under different scene conditions. Thresholding methods determine threshold values using different properties of the image data (Sezgin & Sankur, 2004) which are affected by the change in brightness levels in the camera scene and noise introduced (Yazdi & Bouwmans, 2018). An inappropriate threshold value can change the results of moving object detection significantly. Therefore, it is critical to select an appropriate threshold value. The manual application of a threshold
value is very common but, at the same time, it is challenging to update the threshold value with every scene change using trial and error methods. Automatic thresholding methods determine threshold value using pixel statistics which do not consider the spatial distribution of pixels. Savas et al. used an adaptive thresholding method which employs a counter structure (Savas, et al., 2018). Otsu threshold (Otsu, 1979), adaptive threshold (Bradely & Roth, 2007; Wellner, 1993), iterative Gaussian clustering (Riddler & Calvard, 1978), histogram-based thresholding (Glasbey, 1993), and minimum error thresholding (Kittler & Illingworth, 1986) are some of the commonly used image thresholding techniques described in the next section.

This research article presents a background subtraction technique using Lab colour space. Although Lab colour space has assisted other techniques of moving object detection in removing shadows and noise, direct application of Lab colour space in moving object detection has not been explored to a greater extent. In this study, background and foreground opposite colour pairs of Lab colour space are subtracted (Eqn 3.1) and the differences of red-green and blue-yellow bands are obtained. Utilizing the differences, a novel image thresholding technique is developed in this research article. Difference of opposite colour pairs between foreground and background provide two difference frames which are spatially correlated. An object is detected when the difference of opposite colour pairs is significant. Spatial auto-correlation is used to find highly correlated areas between the difference of red-green and blue-yellow opposite colour pairs. As a result, the basic background subtraction technique is improved by 77% when using Lab colour space and spatial auto-correlation for thresholding in comparison to background subtraction technique when using gray images and existing thresholding methods.
3.2 Review of Existing Image Thresholding Methods

An overview of the most commonly used automatic image thresholding methods used with background subtraction techniques is provided in the following sections.

3.2.1 Otsu Threshold Method

The single-band image, histogram-based Otsu method of image thresholding is widely used in moving object detection techniques. The Otsu method determines a global image threshold value using an image histogram (Otsu, 1979). An image histogram is divided into $N$ number of bins and the number of pixels ($p$) within each bin is counted. It is assumed that a given image histogram consists of two peaks which can be separated by a threshold value among $N$ number of bins. This value is known as an image threshold value ($T$). Using the Otsu method, $T$ is chosen in a way that minimizes the interclass variance of black (pixels values below threshold) and white (pixel values above threshold) clusters of pixels. The within-class variance ($\sigma_{within}^2$) as the weighted sum of variance of each cluster is defined as (Eqn. 3.2):

$$\sigma_{within}^2(T) = n_B(T)\sigma_B^2(T) + n_F(T)\sigma_F^2(T)$$  \hspace{1cm} (3.2)

Where,
\[ P(i) = \frac{p(i)}{\sum_{i=0}^{N-1} p(i)} \]

\[ n_B(T) = \sum_{i=0}^{T-1} P(i) \]

\[ n_F(T) = \sum_{i=T}^{N-1} P(i) \]

\[ \sigma_B^2(T) = \text{Variance of pixels in the background (below threshold)} \]

\[ \sigma_F^2(T) = \text{Variance of pixels in the foreground (above threshold)} \]

\([0, N-1]\) is the range of intensity (or pixel) levels.

The Otsu method finds the pixel value corresponding to the minimum point between two peaks. Any two image histogram distributions are clearly separated if there exists a distinct valley between them. The Otsu method works efficiently under such conditions.

### 3.2.2 Adaptive Image Threshold Method

The image threshold value used in moving object detection needs to be updated continuously due to changing light conditions, image sensor internal noise, reflections from multiple light sources in the camera scene, and pixel values of video frames. To overcome the problem of changing camera scenes, the Adaptive Image Thresholding Method is used which is an extension of Wellner’s method (Bradley & Roth, 2007; Wellner, 1993). In this method, a different threshold value is computed for each pixel of an image. An integral image (also known as a summed-area table) is then computed for each pixel \(I(x,y)\) (Eqn. 3.3).
\[ I(x, y) = f(x, y) + I(x - 1, y) + I(x, y - 1) - I(x - 1, y - 1) \]  \hspace{1cm} (3.3)

Where function, \( f(x, y) \) for any rectangle with upper left corner coordinates \((x_1, y_1)\) and lower-right corner coordinates \((x_2, y_2)\) is computed as (Eqn. 3.4):

\[
\sum_{x=x_1}^{x_2} \sum_{y=y_1}^{y_2} f(x, y) = I(x_2, y_2) - I(x_2, y_1 - 1) - I(x_1 - 1, y_2) - I(x_1 - 1, y_1 - 1)
\]  \hspace{1cm} (3.4)

A moving average of integral image is computed for a window of size \((s \times s)\). If the pixel value of the integral image \( I(x, y) \) is less than the moving average value, the pixel at location \((x, y)\) is set to black; otherwise it is set to white (Eqn 3.5).

\[
I(x, y) = \begin{cases} 
1, & \text{if } I(x, y) \geq \text{moving average } (x, y) \\
0, & \text{otherwise}
\end{cases}
\]  \hspace{1cm} (3.5)

Where,

\[
\text{Moving average } (x, y) = \sum_{x=s}^{x+s} \sum_{y=s}^{y+s} I(x, y) / (s \times s)
\]

3.2.3 Iterative Gaussian Clustering Method

In the Iterative Gaussian clustering method, an arbitrary threshold is selected as the initial threshold based on the assumption that clusters of pixels follow a Gaussian distribution. These pixels are divided into black and white clusters. The Mean of both clusters is calculated and used as a new threshold value. The process is repeated until old and new threshold values are not significantly different (Figure 3.1) (Ridler & Calvard, 1978; Şahan, et al., 2016).
3.2.4 Histogram-Based Thresholding Method

A gray image histogram is divided in two parts in the histogram-based image thresholding method (Glasbey, 1993). The image threshold value depends on the shape of the histogram. The best threshold value is determined only when the histogram contains two clearly separated classes. The optimal threshold value is determined differently using existing methods of histogram-based image thresholding. Histogram shape analysis by measuring concavity, multimodal statistical modeling, and entropy are some of the commonly used methods.
3.2.5 Moving Average Histogram Thresholding

The moving average is used to find the pixel value corresponding to the minimum point in the valley between two peaks of an image histogram. A window of three pixels is used and their average \((p_{n-1}, p_n, p_{n+1})\) is computed. Pixel \(p_n\) is then replaced by the average value (Glasbey, 1993). The averaging process is repeated until the histogram envelope is reduced to an envelope with two peaks. Pixel values corresponding to the two peaks are named \(p_{h1}\) and \(p_{h2}\) and the average of peak pixels \(((p_{h1} + p_{h2})/2)\) is chosen as the threshold value.

3.2.6 Histogram Entropy-Based Image Thresholding

The concept of entropy is originally derived from thermodynamics as a measure of dispersal energy which is used in data transfer over a noisy channel. In image thresholding, background and foreground image pixels are considered as two different image signals. The entropy of each signal is calculated and summed. When the sum has the largest value, the threshold is considered optimal. If \(T\) is the threshold of the gray image, probability distribution of pixels whose value is less than \(T\) is given by Eqn. 3.6 (Kapur, et al., 1985; Carabias, 2012)

\[
\frac{p_0}{p_B}, \frac{p_1}{p_B}, \ldots, \frac{p_T}{p_B}
\]

The probability distribution of pixels greater than \(T\) is given by Eqn. 3.7.

\[
\frac{p_{T+1}}{1-p_B}, \frac{p_{T+2}}{1-p_B}, \ldots, \frac{p_{N-1}}{1-p_B}
\]

Where: \(p_i\) is the probability of occurrence of a gray pixel \(i\) and for an 8-bit image, \(N\) is equal to 256 (or 0 to 255). \(p_B\) is the sum of probabilities of black pixels and \(p_w\) is the sum of probabilities of white pixels (Eqn. 3.8).
\[ p_B = \sum_{i=0}^{T} p_i \text{ and } p_W = 1 - p_B \tag{3.8} \]

Entropy of black and white pixels is computed using Eqn. 3.9 and Eqn. 3.10.

\[ E_B = -\sum_{i=0}^{T} p_i \frac{p_i}{p_B \log(p_i/p_B)} \tag{3.9} \]

\[ E_W = -\sum_{i=T+1}^{N-1} p_i \frac{p_i}{p_W \log(p_i/p_W)} \tag{3.10} \]

For each pixel in an image, black and white pixel entropies are calculated and added. The image threshold is selected such that the sum of entropies \((E_B + E_W)\) is maximized.

### 3.2.7 Minimum Error Thresholding

Kittler and Illingworth (Kittler & Illingworth, 1986) derived image histogram observations from two normal distributions with respective means and standard deviations referred to as \((\mu_1, \sigma_1^2)\) and \((\mu_2, \sigma_2^2)\). These two distributions with weights \(w_1\) and \(w_2\) are separated and the threshold value is determined. The mixture of distributions reflected in the histogram (Carabias, 2012) is given by Eqn 3.11.

\[ f(i) = \frac{w_1}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2}\left(\frac{i - \mu_1}{\sigma_1}\right)^2} + \frac{w_2}{\sqrt{2\pi}\sigma_2} e^{-\frac{1}{2}\left(\frac{i - \mu_2}{\sigma_2}\right)^2} \tag{3.11} \]

Where,

\[ w_1(t) = \sum_{i=0}^{t} P(i) \]

\[ w_2(t) = \sum_{i=t+1}^{N-1} P(i) \]

Each pixel value of the image histogram is considered as a threshold \((t)\) value in sequence. The histogram is divided into two populations using these threshold values. One has pixel values smaller than \(t\) and the other population has pixels larger than \(t\). Weights
and standard deviations of each population are calculated for each threshold value. Later a fitting criterion \( J \) is calculated for each threshold value (Eqn. 3.12).

\[
J(t) = 1 + 2[w_1(t) \log \sigma_1(t) + w_2(t) \log \sigma_2(t)] - 2[w_1(t) \log w_1(t) + w_2(t) \log w_2(t)]
\] (3.12)

The better the model fits the image data, the smaller the value of criterion \( J \). The threshold value \( t \) where \( J \) has the least value is selected as the image threshold.

### 3.3 Methodology

Image thresholding algorithms are used with moving object detection techniques using gray images. A colour and spatial correlation-based image thresholding technique is developed in this paper. RGB colour images are converted to the opposite colour space which is also known as Lab colour space. A high-level process diagram of the complete methodology of moving object detection using Lab colour space and spatial correlation-based image thresholding is shown in Figure 3.2.

![Figure 3.2. Spatial correlation-based image thresholding algorithm](image)

#### 3.3.1 RGB Colour Space to Lab Colour Space Conversion

Most of the still and video cameras capture images in RGB colour space and this is because of semiconductor (material used in camera sensor) spectral characteristics for Red,
Green and Blue lights. Pixel values in the three colour bands of RGB colour space are correlated with each other which makes it irrelevant to use them in moving object detection. The effect of correlation among RGB bands is removed by converting the image from RGB colour space to Lab colour space. Colour and illumination \((L)\) information is separated in the Lab colour space. Also, the coordinates \(a\) and \(b\) represent red-green and blue-yellow opposite colour pairs respectively. The rectangular coordinates \(L, a,\) and \(b\) (Alessi, et al., 2004) are defined as (Eqn. 3.13):

\[
\begin{align*}
L^* &= 116 f(Y/Y_n) - 16 \\
a^* &= 500 [f(X/X_n) - f(Y/Y_n)] \\
b^* &= 200 [f(Y/Y_n) - f(Z/Z_n)]
\end{align*}
\] (3.13)

Where,

\[
f(X/X_n) = (X/X_n)^{1/3} \quad \text{if} \quad (X/X_n) > (24/116)^3
\]

\[
f(X/X_n) = (841/108)(X/X_n) + 16/116 \quad \text{if} \quad (X/X_n) \leq (24/116)^3
\]

And,

\[
f(Y/Y_n) = (Y/Y_n)^{1/3} \quad \text{if} \quad (Y/Y_n) > (24/116)^3
\]

\[
f(Y/Y_n) = (841/108)(Y/Y_n) + 16/116 \quad \text{if} \quad (Y/Y_n) \leq (24/116)^3
\]

And,

\[
f(Z/Z_n) = (Z/Z_n)^{1/3} \quad \text{if} \quad (Z/Z_n) > (24/116)^3
\]

\[
f(Z/Z_n) = (841/108)(Z/Z_n) + 16/116 \quad \text{if} \quad (Z/Z_n) \leq (24/116)^3
\]

Where, \(X, Y,\) and \(Z\) are tristimulus values of the test colour object stimulus and \(X_n, Y_n,\) and \(Z_n\) are tristimulus values of a specified white object colour stimulus. \(X, Y,\) and \(Z\) tristimulus values are obtained from RGB colour values using the following equations.

\[
\begin{align*}
X &= 2.768892R + 1.751748G + 1.130160B \\
Y &= 1.000000R + 4.590700G + 0.060100B \\
Z &= 0 + 0.056508G + 5.594292B
\end{align*}
\] (3.14)
Where coefficients values are obtained using spectral response of standard source and colour matching functions over visible range of electromagnetic spectrum (Alessi, et al., 2004).

3.3.2 Background and Foreground Subtraction

Lab colour bands of background and foreground images are subtracted and the difference bands of $L$, $a$, and $b$ are computed (Eqn 3.15). The difference between background and foreground frames is further processed for moving object detection.

$$
\Delta L = |L_{\text{Background}} - L_{\text{Foreground}}| \\
\Delta a = |a_{\text{Background}} - a_{\text{Foreground}}| \\
\Delta b = |b_{\text{Background}} - b_{\text{Foreground}}| 
$$

3.3.3 Spatial Autocorrelation

With the assumption that the moving object is present in the difference of colour bands (Eqn. 3.15), the physical location of pixels corresponding to the moving object should overlap in $\Delta a$ and $\Delta b$ bands. The spatial correlation between $\Delta a$ and $\Delta b$ is determined. Three different methods of spatial correlation are used to determine image threshold value. These methods are Moran’s Index, Jaccard’s Index, and cross-correlation.

3.3.3.1 Moran’s Index

Local Indicators of Spatial Association (LISA): A Local Indicator of Spatial Association (LISA) satisfies the following two requirements (Anselin, 1995):
1. A LISA for each observation gives an indication of the extent of spatial clustering of similar values around that observation; and
2. The sum of LISAs for all observations is proportional to the global indicator of spatial association.

Local spatial clusters, also referred to as “hot spots”, can be identified as those locations or sets of contiguous locations for which the value of LISA is significant. Similar to the rationale behind significance tests for $G_i$ and $G^*_i$ statistics by Getis and Ord (Getis & Ord, 1992), general LISA can be used as the basis for a test of the null hypothesis of no local spatial association (Ord & Getis, 1995).

**Local Moran:** A local Moran statistic for an observation $i$ is defined as (Anselin, 1995):

$$I_i = z_i \sum_j w_{ij} z_j$$  \hspace{1cm} (3.16)

Where, $z_i$ is the deviation of an attribute for feature $i$ from its mean ($p_i - \bar{P}$) and $w_{ij}$ is the spatial weight between feature $i$ and $j$. Weight between same features is counted as zero ($w_{ii} = 0$).

Moran’s $I$ statistic for spatial autocorrelation is given in Eqn. 3.17:

$$I = \frac{n \sum_i \sum_j w_{ij} z_i z_j}{\sum_i z_i^2}$$  \hspace{1cm} (3.17)

Where, $n$ is the total number of features and $S_0 = \sum_i \sum_j w_{ij}$ is the sum of all weights.

Moran’s $I$ statistics is bounded by -1 and +1. A value of -1 indicates perfect dissimilarity, 0 indicate no auto-correlation and +1 indicate perfect similarity means highly auto-correlated.

**Windowed Operation to Calculate Moran’s Index:** Difference bands of opposite colour pairs (Eqn. 3.15) are divided into small segments using a moving window of size 3x3 (or
a higher odd number). The Moran’s index between window segments of $\Delta a$ and $\Delta b$ is then calculated and assigned to the pixel corresponding to the center of that window. This way a matrix consisting of Moran’s indices is created (Figure 3.3). This process is repeated for different widow sizes and multiple Moran’s index matrices are created. The final Moran’s index matrix (Eqn 3.18) is obtained by summing all index matrices. A high positive index represents a high positive correlation coefficient and a low negative index represents high negative correlation.

$$mI_{x,y} = \sum_{j=w_1}^{w_2} \sum_{w} I_{w_j}$$

(3.18)

Where, $w_1$ is the smaller window (e.g. 3x3) and $w_2$ is the larger window (e.g. 5x5). $I_{w_j}$ is the Moran’s Index between $\Delta a$ and $\Delta b$ for window $w_j$. 

**Figure 3.3.** Spatial correlation (Moran’s Index) calculation between two image bands
3.3.3.2 Jaccard Index

The classical similarity measure between any two sets A and B is given by the Jaccard coefficient which is computed using Eqn. 3.19 (Jaccard, 1912). The similarity between two sets A and B increases with an increase in the Jaccard coefficient (Niwattanakul, et al., 2013).

\[ J(A, B) = \frac{|A \cap B|}{|A \cup B|} \]  (3.19)

The Jaccard distance is calculated between binary sets. Therefore, frame differences \( \Delta a \) and \( \Delta b \) are first binarized using standard deviation values of the sets. If standard deviations of \( \Delta a \) and \( \Delta b \) are \( \sigma_{\Delta a} \) and \( \sigma_{\Delta b} \) respectively, binarized sets of difference bands are calculated as (Eqns. 3.20 and 3.21):

\[
\Delta a^B = \begin{cases} 
1, & \Delta a > \sigma_{\Delta a} \\
0, & \text{otherwise}
\end{cases} \quad (3.20)
\]

\[
\Delta b^B = \begin{cases} 
1, & \Delta b > \sigma_{\Delta b} \\
0, & \text{otherwise}
\end{cases} \quad (3.21)
\]

In order to remove noisy pixels and detect the moving object in a camera scene, Jaccard coefficients are calculated locally using small windows. A sliding window operation is used and the Jaccard coefficient is calculated for each window. For a given window size, the calculated Jaccard coefficient is assigned to the pixel corresponding to the center of the window. At the end, coefficients at each location are added together and final the Jaccard coefficient matrix is computed as given in Eqn. 3.22.
\[ J(\Delta a, \Delta b)_{x,y} = \sum_{j=1}^{w_2} J(\Delta a_{w_j}^R, \Delta b_{w_j}^R) \]

### 3.3.3.3 Cross-Correlation

Spatial cross-correlation is computed between two matrices \( \Delta a \) and \( \Delta b \) locally. If \( \Delta a \) window size is \( M \times N \) and \( \Delta b \) window size is \( m \times n \), 2D cross-correlation of size \( M + m - 1 \) and \( N + n - 1 \) is given as (Proakis & Manolakis, 1996; MathWorks, 2018):

\[
C(x, y) = \sum_{p=0}^{M-1} \sum_{q=0}^{N-1} \Delta a(p, q) \overline{\Delta b(p - x, q - y)}
\]

(3.23)

Where, \(- (m - 1) \leq x \leq M - 1 \) and \(- (n - 1) \leq y \leq N - 1 \) and \( \overline{\Delta b} \) is the complex conjugate of \( \Delta b \). The correlation coefficient for each pixel is computed using windows of different sizes in odd numbers ranging from \( 3 \times 3 \) to \( 13 \times 13 \). For a given window size, two types of window operations are applied, sliding and non-sliding.

**Sliding Window Operation:** In a sliding window operation, the window moves from one pixel location to the next on \( \Delta a \) and \( \Delta b \) bands. The cross-correlation value computed between the two bands is assigned to a location corresponding to the center position of the window.

**Non-Sliding Window Operation:** In a non-sliding window operation, the window jumps from one pixel location to another based on \( \Delta a \) and \( \Delta b \) bands such that the windows do not overlap from one location to another. The computed cross correlation value is assigned to all pixel locations within the window.

Sliding and non-sliding window operations are repeated for windows of different sizes. A single coefficient matrix is generated for each window size. Finally, all matrices are summed together to obtain the final correlation matrix between \( \Delta a \) and \( \Delta b \) (Eqn 3.24).
\[ CC(x, y) = \sum_{j=1}^{w_2} w_j C_{wj} \]

### 3.3.3.4 Random Nature of Noise

Difference frames \( \Delta a \) and \( \Delta b \) contain a partial or fully detected moving object. Depending on the scene, colours, and light conditions, these difference frames also contain noise with random location. It is possible that \( \Delta a \) and \( \Delta b \) may not contain noise together at one place. Therefore, spatial correlation is useful in removing noise. Random noise between two difference bands is ignored by removing negative and low correlated coefficients. Only positive correlation coefficients (in the coefficient matrix) are considered for further processing.

### 3.3.3.5 Coefficient Thresholding

A moving object is detected by calculating image thresholding coefficients using Moran’s Index, the Jaccard Index, and the cross-correlation coefficient. An image threshold is applied to all positive coefficients and a pixel is marked as a moving object only if the correlation coefficient is greater than the threshold value (Eqn. 3.25).

\[
\text{Moving Object Pixel} = \begin{cases} 
\text{Yes}, & \text{if Correlation Coefficient} > Th \\
\text{No}, & \text{Otherwise}
\end{cases}
\]

Due to the different statistics used in different image thresholding methods, the threshold \( Th \) value is selected differently. The criteria to select the threshold values using different spatial correlation coefficients are given in Table 3.1.
Table 3.1. Threshold selection criteria using spatial correlation coefficients

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Threshold ((Th))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross-Correlation (sliding window)</td>
<td>(0.1 \times \text{max (Correlation Coefficient)})</td>
</tr>
<tr>
<td>Cross-Correlation (distinct window)</td>
<td>(0.1 \times \text{max (Correlation Coefficient)})</td>
</tr>
<tr>
<td>Moran’s Index</td>
<td>(0.1 \times \text{max (Moran’s Index)})</td>
</tr>
<tr>
<td>Jaccard Index</td>
<td>(0.9 \times \text{max (Jaccard Index)})</td>
</tr>
</tbody>
</table>

3.4 Results and Discussion

3.4.1 Image Dataset for Moving Object Detection

![Indoor and outdoor camera scenes](image)

Figure 3.4. Indoor and outdoor camera scenes

The image dataset with a single moving object is collected in both indoor and outdoor conditions. For indoor conditions, an object starts moving towards the camera from the far end of a long hallway illuminated by regular fluorescent lights. The lights were adjusted to collect data in two different illumination conditions as High and Medium lights (Figure 3.4a, 3.4b). A parking lot is selected as an outdoor scene where natural sun illumination is the light source. The moving object and camera scenes were not exposed to direct sunlight (Figure 3.4c).
3.4.2 Moving Object

The method developed in this paper for moving object detection using Lab colour space and spatial auto-correlation is evaluated by detecting three different moving objects in indoor and outdoor scene conditions. Data are collected for a single moving object in each scene. The moving objects are distinct in terms of shape and appearance (Figure 3.5).

Figure 3.5. Different moving objects in different camera scenes

The results of moving object detection using the methodology discussed in section 3.3 are shown in Figure 3.6, Figure 3.7 and Figure 3.8. Outputs from the developed methodology are compared with other commonly used background subtraction techniques using different image thresholding methods. Four different methods of thresholding using

Figure 3.6. Moving object detection in indoor scene and medium light condition
spatial auto-correlation are implemented and the moving object is detected using the background subtraction technique. These methods are (1) Cross correlation using sliding window, (2) Cross correlation using distinct window, (3) Moran’s index, and (4) Jaccard coefficient. Six existing methods of image thresholding used to compare outputs are (1) Iterative Gaussian, (2) Entropy, (3) Minimum error thresholding, (4) Moving average, (5) Otsu and (6) Adaptive thresholding.

Among the 6 methods, only Otsu and Adaptive thresholding detected the moving object (Otsu and Adaptive threshold values are computed using Matlab commands graythresh and adaptthresh respectively). Figure 3.6, 3.7, and 3.8 show the results of moving object detection in the medium light indoor scene, high light indoor scene, and outdoor scene respectively. Columns from left to right in each figure show moving object detection results with cross correlation using sliding window, cross correlation using distinct window, Moran’s index, Jaccard coefficient, Otsu thresholding, and Adaptive thresholding, in sequence. It can be observed that moving object detection failed using the Adaptive thresholding technique. Noisy pixels are detected across frames in 8 out of 9 camera scenes.
An object of a darker shade of colour could not be detected (Figure 3.6, row 3) when the test was conducted indoors under medium light conditions. However, other objects of lighter shades were visible under similar light conditions (Figure 3.6, row 1 and 2). The effect of light conditions in a camera scene is demonstrated in Figure 3.6 (row 3) where an object of darker shade was detected in indoor high light conditions. The Object detection rate using Otsu and Adaptive thresholding techniques was lower than techniques using Lab colour space and spatial segmentation. Both Otsu and Adaptive thresholding-based techniques failed to detect the moving object as Otsu thresholding detects partial object and Adaptive thresholding detects background with partial object. Although Otsu thresholding failed to completely detect the object, whenever the contrast between background and foreground was better, Otsu thresholding resulted in better object detection (only upper body of moving object) than when the contrast was poor (Figure 3.6 - Figure 3.8).

Figure 3.8. Moving object detection in outdoor scene and daytime light condition

In the indoor and outdoor, medium and high light conditions, spatial correlation-based image thresholding gave better moving object detection results than Otsu or Adaptive thresholding-based detection. Figure 3.9 shows a larger picture of a detected object (left-
Figure 3.9. Detailed detected object in indoor and outdoor scenes

most object in Figure 3.5) in all scene and light conditions. Each column of Figure 3.9
shows moving object detection results from cross correlation using sliding window, cross

correlation using distinct window, Moran’s index, Jaccard coefficient, and Otsu
thresholding. Spatial correlation-based techniques detected a complete moving object in
comparison to a partial detection using Otsu thresholding. A comparison of moving object
detection results from Figure 3.6 - Figure 3.8 is given in Table 3.2. The last column in the
table gives the total number of camera scenes where a moving object was detected vs not
detected. Among the four spatial correlation-based techniques developed in this paper,
thresholding with cross-correlation using sliding and distinct windows gave better results
than thresholding using Moran’s and Jaccard indexes. Boundary and shape of the object
are also better defined when the object is detected using cross-correlation thresholding.
Moran’s and Jaccard index-based thresholding results are noisier than cross-correlation. Further investigation of results leads to the conclusion that shapes and boundaries of detected object are better defined when the threshold is calculated using cross-correlation distinct window operation (column 2, Figure 3.9). Numerically, out of 9 scenes, the cross-correlation based thresholding method developed by the author in this research detected the target object in 7 scenes (77%) compared to when using the Otsu threshold in 5 scenes (55%) and when using the Adaptive threshold in 1 scene (11%). Four of the existing techniques evaluated in this study failed to detect a moving object in any scene. To conclude, the developed technique shows an improvement of up to 77% in moving object detection results.

Table 3.2. Comparison of moving object detection results using six different techniques of image thresholding
(Y = Yes, Object Detected; N = No, Object Not Detected)

<table>
<thead>
<tr>
<th>Method</th>
<th>Indoor Medium Light</th>
<th>Indoor High Light</th>
<th>Outdoor Daylight</th>
<th>Yes/No</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>O 1</td>
<td>O 2</td>
<td>O 3</td>
<td>O 1</td>
</tr>
<tr>
<td>1. Spatial Correlation using sliding window</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>2. Spatial Correlation using distinct window</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>3. Moran’s Index</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>4. Jaccard Index</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>5. Otsu Threshold</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>6. Adaptive Threshold</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
</tbody>
</table>
3.5 Conclusions

Moving object detection results are affected by the image threshold value used in extracting the moving object from a camera scene. Different image thresholding techniques give different threshold values which result in the detection of moving objects at various levels. Most image thresholding techniques use gray images converted from RGB images. As R, G, and B bands are correlated, moving object detection using them is not feasible. A new background subtraction technique using un-correlated Lab colour space is developed and the image threshold is determined using spatial correlation. A matrix consisting of spatial correlation indices is calculated and the moving object is detected by using the thresholding correlation coefficients with the highest correlation.

The results from the developed techniques are better than the results obtained from Otsu and Adaptive thresholding-based background subtraction techniques. Four spatial correlation determination methods are developed where two of them are based on auto-correlation and the other two are Moran’s and Jaccard indices. The most commonly used Otsu image thresholding method failed to detect the complete object under different light and scene conditions.

Spatial cross-correlation determination using distinct window gave better results in terms of complete detection of the object and its boundary when compared with results from cross-correlation using sliding window, Moran’s Index, and Jaccard Index. In low light and under poor contrast conditions between foreground and background, all the methods under consideration failed to detect the moving object. Techniques developed in this study also detected shadows of the moving object. Results show that poor outputs under low light and poor contrast between foreground and background conditions can be
improved by incorporating contrast enhancement techniques in future. Further, shadows can also be removed by implementing shadow removal techniques. Any additional step can increase the computational cost which will require more research on its applicability and outcomes.

3.6 Acknowledgement

This research was sponsored by the Atlantic Innovation Fund (AIF) of the Atlantic Canada Opportunities Agency (ACOA) with Yun Zhang as the Principal Investigator.

3.7 References


MultiConference of Engineers and Computer Scientists. 1, Mar 13 – 15, 2013, Hong Kong.


Chapter 4: SCORPIQ – A Referenceless Perceptual Image Quality Measurement Technique for Colour Images

Abstract

Quality computation of colour images is important in photography, videography, remote sensing, image fusion and surveillance. Full-Reference and Reduced-Reference image quality measurement techniques require reference images. No-Reference techniques use gray images and are based on image statistics, noise profile, blur and distortion models. However in most cases, the reference image of an input image is not available and gray image quality score does not match the colour perception by human vision system. Literature suggests that human eye perceive the image quality as a combination of colour composition of objects and their spatial distribution. A Spatial Colour Referenceless Perceived Image Quality (SCORPIQ) measurement technique is developed in this paper which is independent of reference image, image statistics and distortion types. Colour segments of an image are spatially analysed with colour information in CIELAB colour space. The SCORPIQ technique gives No-Reference perceptual image quality score to colour images according to human eye perception of colours which can be used to determine the quality of an individual colour image or compare multiple colour images.

4.1 Introduction

Camera technology is growing over time and in its applications. There are different types of cameras used for various applications, such as remote sensing, sports, cinema,

---

professional photography, security, medical imaging, transportation and navigation. Each application utilizes camera images differently. Some of the applications involve visual inspections by the human eyes (or users) and the final product is rated based on the human eye perception. While selecting an image capturing device (camera), users rely on camera specifications for image quality. Under certain circumstances, perceived quality of an image captured by camera can be different and does not match with camera specifications. Each application may require image evaluation differently, which is important and challenging (Figure 4.1). Image quality evaluation techniques are divided into two categories: quantitative (or objective) evaluation uses image pixel values and statistics, and qualitative (or subjective) evaluation uses human inputs in different viewing conditions. Quantitative evaluation may vary across different users, but qualitative evaluation fits the human eye perception model and gives a score irrespective of camera technology or display. Quantitative techniques are easy to implement and most commonly used. Further

Figure 4.1. Three images captured using one camera at one location. Scene condition is different in each image

quantitative techniques are divided into three categories: Full-Reference (FR), Reduced-Reference (RR) and No-Reference (NR) (or referenceless) (Zaric, et al., 2010).

An image consists of one or multiple objects (also known as foreground) and background. Spatial distribution of different objects in an image and their colourfulness
plays an important role in image quality perception. An overall noisy image with recognizable colourful objects (or pattern) distributed across the image may be perceived better by the human eye than a less noisy image with less colourful objects limited to certain areas of the image. By changing the light and scene conditions, one camera can be used to produce images of different qualities which are perceived differently (Figure 4.1), and this is widely used in photography and videography. Therefore, it is difficult to decide the image quality based on camera specifications. In remote sensing, the National Imagery Interpretability Rating Scale (NIIRS) (Irvine, 1997) is used by image analytics to rate satellite images from different sensors. This rating is based on the level of details present in the satellite image. It is possible that there are multiple images with similar ratings, but each image can be perceived differently which may also affect image classification results. Various images (colour) are also used by researchers to produce research outputs. When it is difficult to quantify image quality based on the human eye perception, researchers choose suitable images for their research based on research experience. There is no established criteria about selection of such images which may produce better perceived results out of research. Since one camera can produce differently perceived images, images from different cameras with similar ratings can be perceived differently. FR image quality measurement techniques require reference images all the time; NR image quality measurement techniques (which do not require a reference image) need to be further developed such that any colour image, from any camera, can be assigned a suitable quality score. There are NR image quality techniques which are based on gray images (Zaric, et al., 2010) and, when such techniques are used, quality score does not always match with the human eye perception (or human vision system) (Wang & Bovik, 2006).
Psychological studies have found that the human eye perceives colours using the “opponent theory” of colour vision (Kaiser, 2008). Human eye psychology theory suggests that, while looking at an image, the eye finds the areas with opposite colours (or distant colours) easily and stays focused on those areas most of the time (Wang & Bovik, 2006; Hurvich & Jameson, 1957). An opponent colour pair consists of two of the most distant colours on the colour wheel, and that pair cannot be mixed or seen together by the human eye. Opponent colours are paired as red-green, yellow-blue and black-white. CIELAB colour space (Alessi, et al., 2004) (referred to as Lab colour space in the rest of the article) represents colours according to opponent theory. Each axis in the Lab colour wheel represents an opposite colour pair with opposite colours at the end of axes (Figure 4.2(a)). As shown in Figure 4.2(b), RGB primary colours do not exactly fall on opposite colours, hence the Lab colour space is more suitable for colour analysis than the RGB colour space. Red and Blue are close to each other on the Lab colour wheel, hence less visually

![Figure 4.2. Lab space color wheel (a) Lab color space coordinate system (b) Location of primary RGB colors (Red [255, 0, 0], Green [0, 255, 0], Blue [0, 0, 255] and Yellow [255, 255, 0]) on color wheel (after (Inc., 2018))](image-url)
distinguishable among Red, Green and Blue. Colour distances between two points (or colours) on the Lab colour wheel give the perceived difference of colours by the human eye. Two colours with at least Just Noticeable Distance (JND) can be distinguished by the human eye (Mokrzycki & Total, 2011). The distance in Lab colour space with spatial distribution of colours in image space are useful to determine image quality which follows the human eye perception.

NR (or referenceless) image quality measurements are more useful as they do not require reference images. However, most of the NR techniques are based on gray images and do not include colour information. Colour opponent theory and colour perception by the human eye is used to develop a new Spatial Colour Referenceless Perceived Image Quality (SCORPIQ) measurement technique. The proposed technique converts an input image to Lab colour space where 3D colour distances are measured. The input RGB image is segmented into different colour segments and their spatial distribution is analyzed. Each segment is considered and analyzed with their neighbouring segments. Final image quality is computed as weighted colour distance between segments and their neighbouring segments. The proposed method does not require reference images and it is independent of image capturing, display and processing systems.

4.2 Literature Review

Mean Square Error (MSE) and Peak Signal-to-Noise Ratio (PSNR) are the most commonly used FR image quality measurement techniques. MSE and PSNR are also criticized for not following the human vision in their outputs (Wang & Bovik, 2006; Aja-Fernandez, et al., 2014). These quantitative techniques do not always agree with human
vision perception and give the signal characteristics of the camera employed as an image quality measure. There are also other image quality measurement techniques which incorporate human vision system and use the properties of non-RGB colour spaces (Sheikh & Bovik, 2006; Lee & Plataniotis, 2015; Luo, et al., 2014; Omari, et al., 2015). Since opposite colours cannot be mixed together according to the opponent theory of colour vision (Kaiser, 2008), opposite colours are easily distinguished. With the presence of a distinguishable shape object, natural object or a scene in image with combination of opposite (or distant) colours, the human eye perceives the image better (Alessi, et al., 2004; Wang, et al., 2004; De-Graef, et al., 1990). The larger the colour distance, the easier it is to distinguish colours. The proposed methodology utilizes the colour distance in $Lab$ colour space and colour segment distributions in image space. Final outputs are based on the combination of these two factors.

FR image quality measurement techniques compare an input image with a reference image. Popular FR image quality measurement techniques include, for example, Structural SIMilarity (SSIM) (Wang, et al., 2004), Visual Information Fidelity (VIF) measure (Sheikh & Bovik, 2006), image distortions based on image structure and coherent regions attracting human attention and their interactions (Ghanem, et al., 2008), Feature Similarity Index Measure (FSIM) using phase congruency features obtained from Fourier Transform and Gradient Magnitude (Zhang, et al., 2011), and Perceptual Geometric Distortion (Hsu, et al., 2014). These techniques use different approaches of computing image quality that require a reference image.

On the other hand, RR image quality measurement technique requires a reference image for initial model or training dataset. A model or training dataset should be updated
with time. SSIM is also developed for RR image quality measurement by extracting statistical features from multiscale and multi-orientation reference image (Rehman & Wang, 2012). Quality Aware Clustering (QAC) uses a reference dataset to learn a set of quality-aware centroid (Xue, et al., 2013).

NR image quality measurement techniques do not require a reference image. Most of these techniques are based on image statistics, statistical models, Natural Scene Statistics (NSS) (Saad, et al., 2012), distortion, blocking, noise, blur, entropy, compression or a combination of these features. NSS is used in Distortion Identification-based Image Verity and Integrity Evaluation (DIIVINE) (Moorthy & Bovik, 2011). FR PSNR image quality technique is out-performed by modeling the image naturalness in Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) (Mittal, et al., 2012). JPEG compression degrades the image quality and it is measured by computing no reference PSNR (nPSNR) (Wang, et al., 2012), gradient statistics (Liu, et al., 2010), discrete cosine transform statistics (Saad, et al., 2010) and NSS (Sheikh, et al., 2005). Image blur, noise and blocking are caused by image compression. These are modeled and image quality is computed using Level-of-Detail (Dosselmann & Yang, 2012), training the model using compressed images (Wang, et al., 2002), blur and noise modeling (Choi, et al., 2009; Choi & Lee, 2011; Tang, et al., 2011). Other techniques use entropy (Gabarda & Cristóbal, 2007), Conventional Neural Network (CNN) (Kang, et al., 2014), natural image statistical model (Li, et al., 2016), scale invariant feature transform (Sun, et al., 2014), and visual perception (Fu & Wang, 2016). A detailed review of NR image quality measurement techniques may be found in (Shahid, et al., 2014).
The NR methods listed above use a gray image as an input to compute image quality; however, image perception by the human eyes is impacted by colours (Fairchild, 2005). Perceived quality of a colour image also depends on the objects in an image: in particular, their shape, distribution, association and colours. RGB colour space-based evaluation of colour image quality is computed where colourfulness, sharpness, contrast (Panetta, et al., 2013) and colour pixel difference (Panetta, et al., 2016) are used to determine the quality of the colour image. These measurements are computed in RGB space which does not match with human vision system. Because a Hue polar histogram is used to measure or calculate dominant colour, dispersion and spatial distribution (Ouni, et al., 2012) of colour is not based on the image space. The SCORPIQ measurement technique discussed in this paper is based on colour distances in the Lab colour space and their spatial distribution in image space.

4.3 Methodology

4.3.1 Image Conversion and Processing

Spectral response of semiconductor to red, green and blue electromagnetic spectrum makes it suitable for cameras to capture images in RGB colour space. Semiconductor technology is also used for image display in display systems. Since human eye vision is better represented in the Lab colour space, the input RGB image is converted to an image representing colours in Lab colour space. The rectangular coordinates $L^*$, $a^*$ and $b^*$ are defined as (Alessi, et al., 2004)
\[ L^* = 116 f(Y/Y_n) - 16 \]
\[ a^* = 500 [f(X/X_n) - f(Y/Y_n)] \]
\[ b^* = 200 [f(Y/Y_n) - f(Z/Z_n)] \]  

(4.1)

where
\[
 f(X/X_n) = (X/X_n)^{1/3} \quad \text{if} \quad (X/X_n) > (24/116)^3 \\
 f(X/X_n) = (841/108)(X/X_n) + 16/116 \quad \text{if} \quad (X/X_n) \leq (24/116)^3
\]

and
\[
 f(Y/Y_n) = (Y/Y_n)^{1/3} \quad \text{if} \quad (Y/Y_n) > (24/116)^3 \\
 f(Y/Y_n) = (841/108)(Y/Y_n) + 16/116 \quad \text{if} \quad (Y/Y_n) \leq (24/116)^3
\]

and
\[
 f(Z/Z_n) = (Z/Z_n)^{1/3} \quad \text{if} \quad (Z/Z_n) > (24/116)^3 \\
 f(Z/Z_n) = (841/108)(Z/Z_n) + 16/116 \quad \text{if} \quad (Z/Z_n) \leq (24/116)^3
\]

Where \(X, Y\) and \(Z\) are the tristimulus values of the test colour object stimulus and \(X_n, Y_n\) and \(Z_n\) are the tristimulus values of a specified white object colour stimulus. \(X, Y, Z\) tristimulus values are obtained from RGB colour values using the following equation.

\[
 X = 2.768892R + 1.751748G + 1.130160B \\
 Y = 1.000000R + 4.590700G + 0.060100B \quad (4.2) \\
 Z = 0 + 0.056508G + 5.594292B
\]

Where coefficients values are obtained using spectral response of standard source and caolour matching functuions over visible range of electromagnetic spectrum (Alessi, et al., 2004).
An image and its RGB colours in the Lab colour space are shown in Figure 4.3. The colour wheel in Figure 4.3(b) shows L, a* and b* values of the Lab colour space. For colour visualization, each point on the colour wheel is marked with original colour values from RGB image. Spatial distribution of colours can be seen in Figure 4.3(a) and visual separation of perceived colours can be seen in Figure 4.3(b).

Figure 4.3. A camera produced image as seen by human eyes. (a) RGB image where objects are constructed by spatially arranging RGB colours (Sheikh, et al., 2005), (b) Representation of colours as seen in (a) in three dimensional Lab color space (x: Red (+a) – Green (-a), y: Yellow (+b) – Blue (-b) and z: Black (L = 0) – White (L = 100))

4.3.2 Image Quality Measurement

An image consists of different types of objects. Colours, shape and spatial distribution of objects play an important role in image perception. Areas of uniform colours in an image are separated by image segmentation. Spatial distribution of segments is analyzed to compute final image quality. An overview of image quality measurement methodology is given in Figure 4.4.
A semi-automated image segmentation technique is used to obtain an optimal segmented image. To achieve this, sample colour regions are manually selected on different parts (or objects) of the image, and then regions of matching colours to the sample region are calculated (Figure 4.5) in the Lab colour space. If $L_{ms}$, $a_{ms}$ and $b_{ms}$ are mean Lab colours from sample region, image pixels with matching colours are found as given in Eqn. 4.3.

$$
Seg^R(x, y) = \begin{cases} 
1 & \text{if } dE \leq th \\
0 & \text{if } dE > th
\end{cases}
$$

(4.3)

Where $Seg^R(x, y)$ is segment of pixels belonging to region $R$, $dE$ is 3D colour distance in Lab colour space defined as:

$$
dE = \sqrt{(L(x, y) - L_{ms})^2 + (a(x, y) - a_{ms})^2 + (b(x, y) - b_{ms})^2}
$$

and $th$ is threshold distance determined based on image histograms of sample region. The threshold value is determined using trial and error method and chosen in a way such that small segments are not selected. Any small segments are manually removed.

The process is repeated for each sample region and image segments are found. A group of connected pixels are assigned as to comprise a “segment” and a separate mask is created for each segment.
Figure 4.5 shows the process of sample region selection followed by image segment determination. Unwanted holes in segments are filled using morphological operations. Segment boundary is extracted using morphological operations (erosion) (Eqn. 4.4) and then converted into polygons for further processing. Nested segments are also taken into consideration of image quality measurement.

\[
Seg^B(x,y) = Seg^R(x,y) - Erosion(Seg^R(x,y))
\] (4.4)

4.3.2.2 Neighbourhood Segments

Perceived image quality depends on spatial distribution of objects and their colour distribution. The spatial distribution of segments is analyzed and their neighbouring segments are found. Segments are considered one by one, and nearest segments in all directions are found for each. For a segment under consideration, angular field of view in a certain direction is found by calculating the maximum extent of the neighbouring segment in that direction. The closest segments with complete or partial line of sight in the field of view are considered neighbouring segments (Figure 4.6).
Nested Segments: In certain conditions it is also possible that one or more segments are found inside another segment. If a segment $Seg_a$ contains only one segment $Seg_b$ inside it, there is only one neighbouring segment to $Seg_b$. If a segment $Seg_a$ contains multiple segments inside it, all inside segments are processed together and the sub-segment colour score is measured. Later, this sub-segment colour score is incorporated into the final image quality score.
4.3.2.3 Segment Distribution

The boundary of each segment is divided into sub-segments (Figure 4.6(c)) equal to the number of segments \( n \) in the neighbourhood. Sub-segment boundary length is computed, followed by the weight of each sub-segment \( W_{Sn} \) using the perimeter of the segment under consideration (Eqn. 4.5). This process is repeated for all segments and for nested segments.

\[
\forall \; W_{Sn} = \frac{\text{sub-segment length}}{\text{segment perimeter}} \tag{4.5}
\]

4.3.2.4 Segments colour distance

Image pixels within each segment are found and the mean Lab colour for each segment \((L_s, a_s, b_s)\) is computed. The Euclidean distance between two points in the Lab colour space is known as “colour distance” and its value represents the colour difference perceived by the human eye. For non-saturated colours, a calculated distance of 1 between two colours represents the approximate minimum perceptible difference (Mokrzycki & Total, 2011). The greater the colour difference between two colours, the better the colours are separated by the human eye. For a segment \( S \) under consideration, 3D Lab colour distance is computed between \( S \) and each of \( n \) neighbouring segments \((L_n, a_n, b_n)\) as given in Eqn. 4.6.

\[
\forall \; D_{Sn} = \sqrt{(L_S - L_n)^2 + (a_S - a_n)^2 + (b_S - b_n)^2} \tag{4.6}
\]

4.3.2.5 Segment Quality Score

Segment colour score is determined using sub-segment weights and neighbourhood segment colour distances. If \( W_{sn} \) represents the sub-segment weights of \( n \) sub-segments and
\(D_{sn}\) is the 3D Lab colour distance between segment \(S\) and \(n\) neighbourhood segments, the overall segment (and nested segment) colour quality score \(Q_{seg}\) for that segment is computed using Eqn. 4.7.

\[
\forall \ \forall_{seg} \ Q_{seg} = \sum_{i=1}^{n} W_{si} D_{si}
\]  

(4.7)

4.3.2.6 Final Image Quality

A segment’s quality score depends on sub-segment weights, the number of neighbouring segments and their respective 3D colour distance. If: (1) a segment is surrounded by a larger neighbourhood segment, and (2) the colour difference between the two is larger compared to another segment under similar conditions, then these two segments will be more distinguishable to the human eye. This is because the corresponding weight of a larger neighbouring segment – as calculated using Eqn. 4.5 - will be larger compared to a smaller neighbouring segment, so the corresponding segment colour quality will be larger. Individual segment colour quality scores can also be used to find the most recognizable areas in an image. Segment colour quality scores, including nested segment quality scores, are used, and final image quality \(IQ\) is computed as the sum of all segment colour qualities (Eqn. 4.8).

\[
IQ = \sum_{i=1}^{Seg} Q_i
\]  

(4.8)
4.4 Results and Discussions

Image segmentation based on Lab colour space is sensitive to colour variation; therefore, unwanted noise pixels are also detected and segments are overlapped at the boundaries. The unwanted noise segments are eliminated by filtering out segments with small areas and overlapping regions are merged with segment with similar colours (Figure 4.7).

![Figure 4.7](image1.png)

**Figure 4.7.** Color segments and their boundaries (a) Segment boundaries detected in Lab color space (b) Segment boundaries after removing unwanted small segment and eliminating overlapping areas

Spatial analysis of segments (e.g. Figure 4.6) is done by analyzing each segment one by one. As discussed in section 4.3.2, neighbourhood segments to each segment are found, and their quality scores are computed. Final image quality is computed as discussed in Eqn 4.8. The greater the value of image quality score, the better the image is perceived by the human eye. Since the image quality score out of this method is not scaled to a uniform maximum value, two or more images are compared based on the image quality score. Further, image quality scores from proposed SCORPIQ methodology are evaluated by comparing quality scores with other commonly used NR image quality measurement
techniques. The NR techniques used to compare outputs are BRISQUE (Mittal, et al., 2012), quality evaluation due to JPEG compression (Wang, et al., 2002) (uses 8×8 window), image quality measurement due to blur and noise (Choi, et al., 2009), and no reference nPSNR (Wang, et al., 2012) (uses 8×8 window). These NR quality measurement techniques use gray images as input where quality scores are not sensitive to colour values. Three types of datasets are used to evaluate SCORPIQ image quality measurement technique. Datasets are combinations of colour variant/invariant and spatially variant/invariant images.

4.4.1 Quality of images with varying colours at same pixel location

Colour variant and spatially invariant images are obtained by manipulating lightness and colour values of the colour chart (Figure 4.8(a)) in the Lab colour space. Spatial locations of pixels are unchanged in all images given in Figure 4.8. The quality scores of images using the SCORPIQ methodology and other NR techniques are given in Table 4.1.

As the quality scores from NR methods and the SCORPIQ are not scaled to uniform values, comparison of individual image quality based on quality scores is not feasible. As a result, images are compared as a group. Image quality scores from NR techniques are close to each other, which can be seen from results for images with same lightness (e.g., Figure 4.8(a), (b) and (c)). However, the image quality score from the SCORPIQ methodology is sensitive to colours and gives a different quality score for each image in Figure 4.8. With different lightness and colours, images presented in Figure 4.8 are visually distinguishable. The image quality score calculated using the SCORPIQ methodology supports the human eye perception and gives each image a different score, whereas other
NR image quality measurement techniques are sensitive only to lightness, not to the colours. Quality scores from these NR techniques are distinguishable when there is change in lightness (e.g., Figure 4.8(a), (d) and (g)). The highest quality score from the SCORPIQ methodology is given to Figure 4.8(c) (Table 4.1) which is the most recognizable image among other images.

**Figure 4.8.** Color chart images (RGB) for image quality measurements. Original image (a) is modified in Lab color space. Lightness is compressed to 0.7 and 0.3 of original lightness and colors are stretched to 2 and 3 times of original values respectively.
Table 4.1. Quality score of images presented in Figure 4.8

<table>
<thead>
<tr>
<th>Image</th>
<th>SCORPIQ</th>
<th>BRISQUE</th>
<th>JPEG Compression</th>
<th>Blur and Noise</th>
<th>nPSNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>24.16</td>
<td>37.46</td>
<td>13.16</td>
<td>-2.38</td>
<td>46.54</td>
</tr>
<tr>
<td>(b)</td>
<td>34.97</td>
<td>38.19</td>
<td>13.04</td>
<td>-2.45</td>
<td>46.67</td>
</tr>
<tr>
<td>(c)</td>
<td>45.42</td>
<td>38.97</td>
<td>12.82</td>
<td>-2.51</td>
<td>46.88</td>
</tr>
<tr>
<td>(d)</td>
<td>19.98</td>
<td>46.69</td>
<td>13.55</td>
<td>-1.76</td>
<td>48.10</td>
</tr>
<tr>
<td>(e)</td>
<td>31.42</td>
<td>46.14</td>
<td>13.42</td>
<td>-0.99</td>
<td>48.30</td>
</tr>
<tr>
<td>(f)</td>
<td>39.09</td>
<td>46.78</td>
<td>13.14</td>
<td>-1.03</td>
<td>48.36</td>
</tr>
<tr>
<td>(g)</td>
<td>15.41</td>
<td>56.63</td>
<td>14.24</td>
<td>-0.90</td>
<td>51.20</td>
</tr>
<tr>
<td>(h)</td>
<td>23.34</td>
<td>54.73</td>
<td>13.93</td>
<td>-0.91</td>
<td>51.32</td>
</tr>
<tr>
<td>(i)</td>
<td>28.61</td>
<td>54.12</td>
<td>13.49</td>
<td>-0.97</td>
<td>51.14</td>
</tr>
</tbody>
</table>

4.4.2 Quality of images with similar colours at different locations

NR image quality measurements are independent of reference images. Therefore, individual images with different scenes can be compared using NR image quality measurement techniques. Most of the NR image quality measurement techniques are insensitive to spatial distribution of pixels (or objects) and colours. As a result, a similar quality score can be obtained by differently perceived images. To obtain colour invariant and spatially variant images, the colour chart (Figure 4.8(a)) is manipulated by exchanging

![Figure 4.8](image)

Figure 4.9. Image generated after changing spatial locations of color segments on original color chart (Figure 4.8(a))

the positions of colour segments. Three new images as shown in Figure 4.9 are generated and their quality scores are compared (Table 4.2). Overall, statistics, colours and lightness
of newly generated images are very close to each other. As noticed, quality scores from NR techniques are very close and they fail to distinguish three completely different images. By contrast, the scores from the SCORPIQ methodology give a different score to each image in Figure 4.9 where the most recognizable image (Figure 4.9(c)) scores maximum quality score.

Table 4.2. Quality score of images in Figure 4.9

<table>
<thead>
<tr>
<th>Image</th>
<th>SCORPIQ</th>
<th>BRISQUE</th>
<th>JPEG Compression</th>
<th>Blur and Noise</th>
<th>nPSNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>14.74</td>
<td>83.88</td>
<td>11.96</td>
<td>-4.10</td>
<td>46.89</td>
</tr>
<tr>
<td>(b)</td>
<td>16.21</td>
<td>83.58</td>
<td>12.31</td>
<td>-4.06</td>
<td>46.99</td>
</tr>
<tr>
<td>(c)</td>
<td>29.99</td>
<td>83.40</td>
<td>11.07</td>
<td>-4.02</td>
<td>46.83</td>
</tr>
</tbody>
</table>

4.4.3 Colour and Spatially Variant Image Quality

Eleven (11) non-distorted images which are colour and spatially variant are selected from the LIVE database (Sheikh, et al., 2005) and shown in Figure 4.10. The LIVE database images are available with subjective quality scores. Image quality score using SCORPIQ methodology is computed and compared with NR image quality scores and mean quality score from LIVE database. The respective individual quality scores for images shown in Figure 4.10 are given in Table 4.3.
As given in Table 4.3 and comparing two or more images together in Figure 4.10, NR image quality scores neither follow the pattern of LIVE database scores nor SCORPIQ quality scores. A disagreement between LIVE database and SCORPIQ scores is also seen in Table 4.3 except Figure 4.10(k), which is assigned the minimum score using both methods. By comparing 11 images as a group, the SCORPIQ quality score is more spread (score from 13.1 to 39.43 with standard deviation of 8.02) compared to LIVE database score (score from 77.35 to 85.64 with standard deviation of 2.65). The LIVE database score is obtained from ratings from different viewers (Sheikh, et al., 2005) and the quality scores are affected by objects present in the scene and viewer’s interest. An image of parrots covering the major area in image with bokeh effect (object in foreground is in camera focus and background is out of focus or vice-versa) (Figure 4.10(e)) is scored with maximum score followed by an image with airplane in LIVE database. An image with most vibrant
and distinguishable colours Figure 4.10(a) is scored 1st SCORPIQ and 5th in LIVE database. This might be because of object size, distinguishability of objects and interest of viewers.

### Table 4.3. Quality scores of images in Figure 4.10

<table>
<thead>
<tr>
<th>Image</th>
<th>SCORPIQ</th>
<th>LIVE</th>
<th>BRISQUE</th>
<th>JPEG</th>
<th>Blur and Noise</th>
<th>nPSNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>39.43</td>
<td>78.36</td>
<td>17.25</td>
<td>9.12</td>
<td>-3.11</td>
<td>38.08</td>
</tr>
<tr>
<td>(b)</td>
<td>38.86</td>
<td>81.06</td>
<td>2.44</td>
<td>10.05</td>
<td>-3.53</td>
<td>40.12</td>
</tr>
<tr>
<td>(c)</td>
<td>37.69</td>
<td>80.74</td>
<td>19.50</td>
<td>9.99</td>
<td>-2.37</td>
<td>39.92</td>
</tr>
<tr>
<td>(d)</td>
<td>35.31</td>
<td>77.95</td>
<td>10.15</td>
<td>11.36</td>
<td>-2.35</td>
<td>42.38</td>
</tr>
<tr>
<td>(e)</td>
<td>30.80</td>
<td>85.64</td>
<td>-12.36</td>
<td>11.07</td>
<td>-2.39</td>
<td>43.99</td>
</tr>
<tr>
<td>(f)</td>
<td>30.34</td>
<td>83.43</td>
<td>-2.03</td>
<td>10.69</td>
<td>-2.55</td>
<td>42.51</td>
</tr>
<tr>
<td>(g)</td>
<td>28.20</td>
<td>77.93</td>
<td>4.74</td>
<td>9.16</td>
<td>-2.91</td>
<td>38.49</td>
</tr>
<tr>
<td>(h)</td>
<td>27.19</td>
<td>78.07</td>
<td>4.11</td>
<td>11.28</td>
<td>-2.87</td>
<td>43.52</td>
</tr>
<tr>
<td>(i)</td>
<td>24.14</td>
<td>79.77</td>
<td>7.39</td>
<td>10.22</td>
<td>-3.24</td>
<td>40.54</td>
</tr>
<tr>
<td>(j)</td>
<td>22.40</td>
<td>78.26</td>
<td>4.25</td>
<td>9.89</td>
<td>-2.35</td>
<td>40.23</td>
</tr>
<tr>
<td>(k)</td>
<td>13.10</td>
<td>77.35</td>
<td>4.78</td>
<td>10.69</td>
<td>-2.54</td>
<td>42.14</td>
</tr>
<tr>
<td>STD</td>
<td>8.02</td>
<td>2.65</td>
<td>8.66</td>
<td>0.78</td>
<td>0.41</td>
<td>1.95</td>
</tr>
</tbody>
</table>

### 4.5 Conclusions

A NR colour-based image quality measurement technique has been discussed. This methodology does not require a reference image. Image quality is computed based on spatial distribution of colour segments and their colour distances from neighborhood segments in the Lab colour space. As discussed in the literature, the human eye perceives opposite (or distant) colours better than nearby colours on the Lab colour wheel. The SCORPIQ methodology gives a quality score which is higher for the images with distant colours, and colour segments are distributed across the image.
As the image quality scores from the SCORPIQ methodology are not scaled to uniform values, quality scores from the commonly used NR techniques are compared as a group. Most of the NR techniques are insensitive to colours and based on modeling of image statistics, noise, blur and other parameters; the SCORPIQ methodology uses colour images as input and considers their spatial distribution. As a result, SCORPIQ methodology gives better image quality score which is in accordance to the human eye perception. Individual images from different or same sources might be scored equally using NR techniques. Using colour information and their spatial distribution, the proposed technique scores them differently.

Further, LIVE database images with different scenes are selected and their quality scores are computed. Quality scores from existing NR techniques are very close to each other, meaning images have similar quality which do not follow quality scored using SCORPIQ methodology. The SCORPIQ methodology gives the maximum score to the most distinguishable image without the requirement of reference image. Since reference image is not available for all the images, NR methods of image quality are critical and SCORPIQ methodology is useful to determine image quality according human eye perception.

SCORPIQ quality scores are also compared with subjective quality scores available with LIVE database which does not follow the SCORPIQ quality score pattern for image set except same image is scored with least quality score using both methods. While LIVE database quality scores also include human interest subject to the objects present in an image, SCORPIQ method scores the images based on colour values and their spatial
distribution in an image. Also, variance of image quality scores using SCORPIQ methodology is better than LIVE database scores.

Processing steps involved in calculating image quality scores include semi-automatic segmentation and do not consider the image noise present in segmented areas. Future research can include better automatic image segmentation techniques and parameters related to noise, type of object, and their sizes. The SCORPIQ measurement technique can also be used with other quality techniques (FR, RR and NR) when key parameters (e.g. blur, noise, statistics etc.) are used in evaluation of different images. Images of the same scene taken from different cameras can also be compared such that a better image (or camera) which is better perceived by the human eye can be found..

4.6 Acknowledgements

This research was sponsored by the Atlantic Innovation Fund (AIF) of the Atlantic Canada Opportunities Agency (ACOA) with Yun Zhang as the Principal Investigator.

4.7 References


Chapter 5: Summary and Conclusions

This chapter presents a summary of research conducted for this dissertation. Outlines of the research from chapters 2 to 4 as well as the contributions of this study are also presented in this chapter. At the end, some suggestions for future work are provided.

5.1 Summary of Research

In this dissertation, the potential of still and video camera outputs is explored for detecting moving objects and measuring their qualities. Chapter 2 and Chapter 3 focus on improving frame differencing and background subtraction techniques of moving object detection using footstep sound and lab colour space respectively. Chapter 4 focuses on the quantification of camera output quality measurements incorporating human eye vision system. A summary of research work performed from Chapter 2 to Chapter 4 is discussed as follows.

5.1.1 Summary of Chapter 2

This chapter investigated frame-based moving object detection techniques. Frame-based techniques use two or more consecutive frames to detect moving objects. These techniques are computationally fast but only detect boundaries of the moving objects. These techniques fail to detect or partially detect an object when it is moving slowly or the image frame is noisy. As frame differencing technique involves simple steps such as the calculation of image difference and thresholding, it is computationally less expensive than other frame-based techniques utilizing Optical Flow and Wavelet Transform. Frame
differencing technique based moving object detection for a single moving object is improved by using footstep sound when the former fails to detect the moving object (a human), partially detects the object, or when the frame is noisy. Detection of footstep sound makes it certain that the object is present in the camera scene. Start and end timestamps of each footstep sound are determined and analyzed for a missing/partially detected object and noisy outputs. An object is modelled as an ellipse which is fitted using the least squares technique on pixels detected as a moving object. A missing/partially detected object in a frame is modelled as a complete object using the ellipses from neighbourhood frames. A spatial segmentation technique is used to remove noise from noisy frames. The methodology discussed in this chapter uses a camera with an inbuilt microphone and the application of footstep sound depend on it being generated from the on-scene moving object. Application of footstep sound has the limitation that the sound is generated from the object present in camera scene. In order to ascertain the source of sound as the moving object, videos were captured under indoor conditions where other moving objects were not present. Results from four different indoor videos show that the object detection rate from frames falling under footstep sound is improved by 52% on using the Frame differencing technique. Object detection rates are also improved by 45% and 31% using Wavelet Transform and Optical Flow frame-based moving object detection techniques respectively.

5.1.2 Summary of Chapter 3

This chapter investigated and improved outputs from most commonly used background subtraction techniques of moving object detection. Most background-based moving object detection techniques use gray images to detect moving objects. However, detection outputs
are affected by the colours of moving objects and the colour contrast between foreground and background. Application of RGB bands is not feasible because of band correlation. Selection of threshold value also plays a critical role in moving object detection. Different image thresholding techniques give different threshold values which result in different levels of moving object detection. A new background subtraction technique using uncorrelated Lab colour space is developed and the image threshold is determined using spatial distribution of pixels. A matrix consisting of spatial correlation indices between the background-foreground differences of opposite colour pairs is calculated and the moving object is detected by thresholding correlation coefficients. Results from the developed techniques are better than the results obtained from Otsu and Adaptive thresholding-based background subtraction techniques. Four spatial correlation determination methods are developed, two of which are based on auto-correlation and the other two on Moran’s and Jaccard indices.

5.1.3 Summary of Chapter 4

This chapter described the methodologies for quality computation of camera outputs without the use of a reference image. No-Reference image quality measurement techniques do not require reference images, but these techniques evaluate image quality based on gray images and pixel statistics. Colour information, which plays an important role in image quality measurement by human eyes, is lost during RGB to gray conversion. In addition, pixel statistics do not have information about spatial distribution of image pixels. A new SCORPIQ, No-Reference colour-based image quality measurement technique is developed in this chapter. Image quality is computed based on spatial distribution of colour segments.
and their colour distances from neighborhood colour segments in the Lab colour space. Results from the developed methodology are compared with other commonly used No-Reference image quality measurement techniques. After visually inspecting results from the developed method, it is noticed that quality scores from SCORPIQ methodology gives a quality score which is according to the spatial distribution of colours and their distances in the Lab colour space. Human eyes distinguish colours according to their distances in the Lab colour space. Images of different scenes from the LIVE database are selected and quality scores are computed for the evaluation of the developed methodology. Existing NR techniques give similar quality scores to visually distinct images whereas SCORPIQ gives different quality scores to different images. SCORPIQ quality scores are also compared with subjective quality scores available from the LIVE database which does not follow the SCORPIQ quality score pattern for image set except the same image is scored with least quality score using both methods. While LIVE database quality scores account for human interest in the objects present in an image, SCORPIQ scores the images based on colour values and their spatial distribution in an image. Variance of image quality scores calculated using SCORPIQ is higher than LIVE database scores.

5.2 Research Achievements

Based on the three main chapters (2 – 4) of this dissertation, a summary of overall contributions is presented as follows:
1.1.1 Improved Frame Differencing moving object detection technique

Experiments conducted in this research show that the moving object (a single human subject) detection rate is improved with the application of footstep sound produced by the object being detected. The frame differencing technique witnessed greater improvement than other frame-based techniques using Wavelet Transform and Optical Flow. The frame differencing technique is less likely to be used for moving object detection due to inefficient results. With improved outputs in indoor conditions, frame differencing technique has created an option to use it for fast processing and reliable outputs.

1.1.2 Improved Background Subtraction moving object detection technique

The technique developed in this research detects the moving object using Lab colour space. Application of opposite colour pairs of Lab colour space ignores the effect of light change which is one of the major reasons responsible for the failure of moving object detection techniques. Inappropriate selection of threshold values is another reason background subtraction techniques either fail to detect or partially detect moving objects. The proposed methodology uses non-correlated opposite colour pairs for moving object detection and a spatial autocorrelation-based segmentation process to extract moving objects. Noisy and false pixels are ignored with the help of spatial autocorrelation used in spatial segmentation. Experimental results show that the proposed technique gives better results under different light and scene conditions than the existing techniques using most commonly applicable thresholding methodologies.
1.1.3 Development of a new No-Reference image quality measurement technique

A novel method is developed to evaluate quality of images and videos out of still and video cameras. The outcome of this research gives a quantitative quality score based on the human eye perception of colours. Spatial distribution of colours and their visual distinguishability are used in computing the final quality score. Compared to the existing Full-Reference and Reduced-Reference techniques, the developed technique does not require a reference image. The existing No-Reference techniques compute image quality based on gray bands which do not consider human eye vision. In comparison, the developed technique uses colour bands and psychological behavior involved in colour perception by humans.

5.3 Suggestions for future work

The proposed technique to improve frame differencing moving object detection used footstep sound obtained from camera inbuilt microphones. The methodology is limited to the detection of footsteps which are generated from the moving object present in the camera scene. Accuracy and certainty of the presence of a moving object in a camera scene can be further improved by using advanced directional microphones, which can record sound only from the camera scene. Relation between camera’s field of view and spatial coverage of microphone needs further investigation. The proposed technique can be further improved for the detection of multiple moving objects. This can be achieved by analyzing higher order MFCC coefficients.

In this research, the object detection rate using the background subtraction technique is improved but the shadows of moving objects are also detected in the process. Contrast
enhancement for frames with poor foreground-background contrast and implementation of shadow removal algorithms can further improve the accuracy of moving object detection outputs.

The SCORPIQ measurement technique of image quality measurement does not require a reference image and can also be used with other image quality measurement techniques (FR, RR and NR) where key parameters such as blur, noise, sharpness, and image statistics are used in the evaluation of the quality of different images. Further, parameters related to the size of objects/segments and their types can also be incorporated in SCORPIQ. Inclusion of such parameters will eliminate the effect of specific interest in objects in evaluating image quality by human eyes.
Curriculum Vitae

Candidate’s full name: Aditya Roshan

Universities attended:

2012 - Present PhD Candidate, University of New Brunswick Fredericton, Canada

2008, MTech, Civil Engineering, Indian Institute of Technology Kanpur, India

2005, MSc, Physics, Banaras Hindu University Varanasi, India

2003, BSc, Mathematics and Physics, Chhatrapati Shahu Ji Maharaj University Kanpur, India

Publications:


