Video Content Analysis for Intelligent Forensics

by

Muhammad Fraz

A Doctoral Thesis

Submitted in partial fulfilment
of the requirements for the award of

Doctor of Philosophy
of
Loughborough University

27th November 2014

Abstract

The networks of surveillance cameras installed in public places and private territories continuously record video data with the aim of detecting and preventing unlawful activities. This enhances the importance of video content analysis applications, either for real time (i.e. analytic) or post-event (i.e. forensic) analysis. In this thesis, the primary focus is on four key aspects of video content analysis, namely; 1. Moving object detection and recognition, 2. Correction of colours in the video frames and recognition of colours of moving objects, 3. Make and model recognition of vehicles and identification of their type, 4. Detection and recognition of text information in outdoor scenes.

To address the first issue, a framework is presented in the first part of the thesis that efficiently detects and recognizes moving objects in videos. The framework targets the problem of object detection in the presence of complex background. The object detection part of the framework relies on background modelling technique and a novel post processing step where the contours of the foreground regions (i.e. moving object) are refined by the classification of edge segments as belonging either to the background or to the foreground region. Further, a novel feature descriptor is devised for the classification of moving objects into humans, vehicles and background. The proposed feature descriptor captures the texture information present in the silhouette of foreground objects.

To address the second issue, a framework for the correction and recognition of true colours of objects in videos is presented with novel noise reduction, colour enhancement and colour recognition stages. The colour recognition stage makes use of temporal information to reliably recognize the true colours of moving objects in multiple frames. The proposed framework is specifically designed to perform robustly on videos that have poor quality because of surrounding illumination, camera sensor imperfection and artefacts due to high compression.

In the third part of the thesis, a framework for vehicle make and model recognition and type identification is presented. As a part of this work, a novel feature representation technique for distinctive representation of vehicle images has emerged. The feature representation technique uses dense feature description and mid-level feature encoding scheme to capture the texture in the frontal view of the vehicles. The proposed method is insensitive to minor in-plane rotation and skew within the image. The capability of the proposed framework can be enhanced to any number of vehicle classes without re-training. Another important contribution of this work is the publication of a comprehensive up to date dataset of vehicle images to support future research in this domain.

The problem of text detection and recognition in images is addressed in the last part of the thesis. A novel technique is proposed that exploits the colour information in the image for the identification of text regions. Apart from detection, the colour information is also used to segment characters from the words. The recognition of identified characters is performed using shape features and supervised learning. Finally, a lexicon based alignment procedure is adopted to finalize the recognition of strings present in word images.

Extensive experiments have been conducted on benchmark datasets to analyse the performance of proposed algorithms. The results show that the proposed moving object detection and recognition technique superseded well-know baseline techniques. The proposed framework for the correction and recognition of object colours in video frames achieved all the aforementioned goals. The performance analysis of the vehicle make and model recognition framework on multiple datasets has shown the strength and reliability of the technique when used within various scenarios. Finally, the experimental results for the text detection and recognition framework on benchmark datasets have revealed the potential of the proposed scheme for accurate detection and recognition of text in the wild.

Acknowledgements

First praise is to Allah, the Almighty on whom ultimately we depend for sustenance and guidance. Second, my deepest appreciation goes to my supervisor Prof. Eran A. Edirisinghe for his guidance and support through the long years of research. His timely and efficient contribution, continuous encouragement and trust in my ideas helped me shape this thesis into its final form. In addition, a thank you to Dr. Saquib Sarfraz; who has the attitude and substance of a genius. His valuable advice and technical support at each stage helped me follow the right direction in my research.

I would like to acknowledge the department of Computer Science, Loughborough University for the financial support provided through my research scholarship. My sincere thanks go to Dr. Gerald Schafer and Dr. Ana Salagean for their annual review on my research progress and advice given. I would like to extend my gratitude to the academic and secretarial staff of the department of Computer Science for their kind assistance. My kind regards to the Visimetrics Ltd. for getting me access to the Loughborough CCTV surveillance centre and the staff there for their help and patience in providing the necessary video data that aided the research work presented in this thesis.

My heartiest thanks go to all the colleagues of my research group. Along with them, I thank all my relatives and friends in my longing home country.

A special thanks to my beloved parents who with all their means and sacrifices have provided me with endless love and all necessities. Furthermore, I am grateful to my adorable sister and brother who with their affection and laughter always helped me to look at the bright side of life.

It is my privilege to extend my deep appreciation and sincere gratitude to my beloved wife Kiran for her immense support, endless patience and all the sacrifices that she made on my behalf during these years. Lastly and most importantly, I'd like to express my endless and unconditional love for my daughter Areesha who became a source of inspiration, positive energy and strength for me through difficult times.

Muhammad Fraz, November 2014

Contents

List of Figures

List of Tables

List of Abbreviations

BF	Bilateral Filter
CC	Connected Components
CD	Cosine Distance
CS	Cosine Similarity
CSLBP	Centre Symmetric Local Binary Patterns
CT	Contourlet Transform
DCD	Dominant Colour Descritpor
DFB	Directional Filter Bank
ED	Euclidean Distance
EM	Expectation Maximisation
FNR	False Negative Rate
FPPW	Flase Positives Per Window
FPR	False Positive Rate
FPPW	False Positives Per Window
FV	Fisher Vector
GMM	Gaussian Mixture Model
HOG	Histogram of Oriented Gradients
LBP	Local Binary Patterns
LESH	Local Energy based Shape Histogram
LP	Laplacian Pyramid
MF	Median Filtering
MLR	Mid Level Representation
MoG	Mixture of Gaussians
MSER	Maximally Stable Extremal Region

MVQ	Minimum Variance Quantization
OCR	Optical Character Recognition
PCA	Principal Component Analysis
ROC	Receiver Operating Characteristics
RoI	Region of Interest
SIFT	Scale Invariant Feature Transform
SVM	Support Vector Machine
SWT	Stroke Width Transform
TPR	True Positive Rate
TNR	True Negative Rate
UQ	Uniform Quantization
VMMR	Vehicle Make and Model Recognition

Chapter 1

Introduction

The use of Closed-Circuit Television (CCTV) cameras for the monitoring of important public and private places has significantly increased during the last decade. Networks of CCTV cameras are in operation to assist security agencies in keeping an eye on suspicious individuals and events. This has increased the necessity for an effective mechanism to manage the growing collection of video data. It is a tedious task to annotate the large amount of video data by manual assignment of meaningful annotations to enable rapid search in case of an event. In this context, at the most basic level, an event of interest is described either as a vehicle or a person passing by a camera.

Humans and vehicles are the two most important subjects for monitoring in any typical video forensics application. They are usually described in terms of appearance, size, behaviour/action (in case of humans) and make, model and type (in case of vehicles). For example, it may be required to identify a person wearing a specific coloured top (appearance) with certain height (size) running (action) through the scene. It may be needed to identify a vehicle with particular type, make, model and colour. Similarly, further detailed annotations such as “a person wearing a short-sleeved, white and blue top that includes text (or a logo)” and black trousers etc. may also be of interest in some applications. Annotation of human and vehicle objects at different levels of detail can help in forensic applications that involve the search for people or vehicles with a known description as seen by a witness.

The work presented in this thesis addressed various challenges involved in the automatic annotation of surveillance videos. The work focuses initially on the classification of foreground moving objects into humans and vehicles. Next, the subsequent detailed annotation of these objects is performed based on a number of appearance measures.

In the case of humans, these measures include:

1. Colour of clothes on different parts of the human body.
2. Text information present in the clothes.

In case of vehicles, these measures include:

1. Vehicle make and model recognition and type identification.
2. Dominant colour of the vehicle.
3. Text information present on the vehicles.

1.1 Research Motivation

The accurate extraction of appearance information in surveillance videos is an extremely challenging task. Typically these videos are recorded at low resolution, low frame rate and their quality is affected by surrounding illumination, camera calibration and compression mechanisms. No standard has been defined as yet about the information present in a personal description for video forensic applications. In most cases, the description mimics a witness's statement and what the witness remembers. Therefore, a limited amount of research exists in the application of computer vision and pattern recognition in surveillance video forensics. Apart from that, mostly the research ([28, 85, 3, 5]) in video analysis considers good quality videos. Therefore, the algorithms developed for good quality videos often fail to perform accurately in low quality videos such as CCTV feeds.

Figure 1.1 shows the block diagram of a video annotation system particularly taking into account the moving targets. The capability of a video annotation system can be extended to any level of detail but the scope of the work presented in this thesis is limited to four major areas of practical importance:

1. Human and Vehicle detection and recognition.
2. Correction and extraction of colours in humans and vehicles.
3. Make and Model recognition of vehicles.
4. Text detection and recognition in outdoor scene.

The motivation for the research covering the context of the investigation presented in this thesis came from a comprehensive literature review and related mathematical and technical areas and the subsequent investigation of shortcomings of the existing algorithms. These deficiencies of the state-of-the-art technology provide the basic research motivation to seek solutions that utilize different approaches to develop complete, end-to-end solutions for the identified problems in the annotation of surveillance videos.

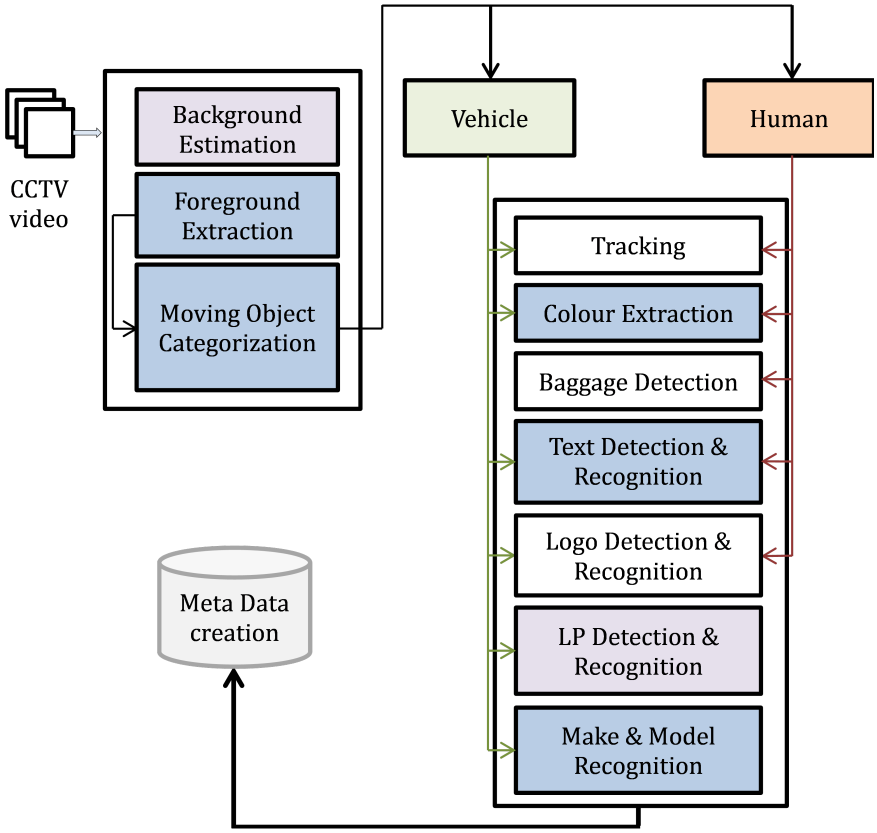


Figure 1.1: The block diagram of a video annotation system depicting key stages. The problems that have been addressed in this thesis, their blocks are highlighted in blue colour.

1.2 Aims and Objectives

The aim of the research work presented in this thesis is the design and development of efficient algorithms for the detection of moving objects and recognition of visual attributes associated with those objects for the purpose of annotating surveillance videos. Detection of humans and vehicles, correction and recognition of colour of human clothes and vehicles, make and model recognition of vehicles and detection and recognition of text present in the outdoor scene are a few of many algorithms that are robust to a number of practical challenges that often limits the present state-of-the-art algorithms being used in practice.

The specific research objectives are listed below:

- To carry out detailed literature review in the areas of research that have been addressed in this thesis and to analyse the challenges that the algorithms may face when implemented and used in practical CCTV systems.

- To develop a robust feature representation technique for the classification of objects particularly humans & vehicles and that can perform better in challenging scenarios as compared to existing algorithms.

- To improve an existing state-of-the-art algorithm in foreground segmentation for shadow removal and accurate contour extraction in order to aid the moving object recognition pipeline.
- To develop a pipeline for correction and recognition of colours of objects present in CCTV videos using existing colour consistency and pixel clustering techniques. The practical challenges faced by existing frameworks especially in the case of surveillance videos have been specifically addressed.
- To develop a novel algorithm for vehicle make and model recognition to perform robustly on a dataset containing large number of vehicle classes. The focus of the developed framework is to achieve high recognition performance on surveillance videos captured under non-controlled conditions.
- To develop a fast and efficient framework for the accurate detection and recognition of text information by exploiting colour information. The devised framework needs to be robust in scenarios where existing state-of-the-art techniques fail.
- To identify the limitations of the proposed algorithms and outline the future directions of research.

1.3 Contribution of the Research

The research work carried out in meeting the above aims and objectives has resulted in a number of original contributions to video analytic and forensic application domains. The key contributions of this thesis are outlined in each subsection below.

1.3.1 Improved Foreground Segmentation Technique

The first step in any video analysis system is the detection and accurate classification of the objects. The presence of background noise and object shadows causes wrong foreground segmentation output. This thesis improves an existing foreground extraction technique to perform robustly in the presence of shadows and cluttered background on low quality CCTV videos.

1.3.2 Contourlet Transform based Centre Symmetric Local Binary Patterns (CCSLBP) Descriptor

The thesis contributes towards the development of a novel feature descriptor for the classification of objects. The proposed feature descriptor enhances the discriminative patterns in the contours of the objects using Contourlet Transform (CT) and then captures them using Local Binary Patterns (LBP). The proposed feature descriptor is mainly used to classify the foreground objects in to humans, vehicles and noise. The proposed technique overcomes the disadvantages of the well-known Histogram of Oriented Gradients (HOG) features for pedestrian detection.

1.3.3 Colour Correction and Recognition Pipeline

An efficient colour correction and recognition pipeline in proposed to extract the dominant colour of moving objects. The proposed pipeline exploits the existing colour constancy algorithm along with a novel adaptive thresholding technique to enhance the true colours of the objects, while cancelling the effect of surrounding illumination. Colour is an important visual feature of objects and can be used for annotation and retrieval of objects in videos. The task of accurate colour recognition of objects becomes more challenging in case of surveillance videos. This is due to the fact that these videos are recorded at low resolution and compressed to reduce the memory requirement for their storage. These factors introduce a high level of noise that affects the clarity and true colours of objects. The proposed framework exploits temporal information by using pixel information from multiple consecutive frames to recognize the dominant colour of objects.

1.3.4 Mid-Level Feature Encoding based Texture Representation

A novel feature representation technique is proposed in the context of vehicle make and model recognition. The proposed feature representation has exhibited the strong capability to capture the textured pattern present in the frontal view of the vehicles. The proposed feature representation scheme computes the shape description around the stable high frequency points with in the image and encodes the features in to another detailed representation through mid-level feature encoding scheme. A single encoded representation is computed for each stable point within the image and the collection of all these representations is used as a vocabulary to represent that particular pattern. The proposed representation performed efficiently in capturing the slightest variation of shapes between two different models of the vehicles.

1.3.5 Vehicle Make and Model Recognition

A further contribution of this thesis is the design, implementation and testing of a novel Vehicle Make and Model Recognition (VMMR) system. The information about a vehicle's make and model can serve as an important parameter for video annotation and retrieval. In comparison to existing VMMR techniques, the proposed system achieves better recognition accuracy on surveillance videos. Another important contribution of the work is the collection of an image dataset and two video datasets. These datasets have been collected to provide a standard benchmark for vehicle make and model research. The proposed system also use a novel image alignment technique to eliminate angular variation in the vehicle images for better recognition performance.

1.3.6 Text Detection and Recognition

Another key contribution of this thesis is the design, implementation and testing of an efficient text detection and recognition system in outdoor scenes. The proposed framework exploits the colour information to detect the potential text regions within the images. A novel character identification technique is also developed to enhance the functionality of Bilateral Regression (BR) [42] based character identification. The proposed technique is extensively tested and experimental results show that it supersedes the original method. The proposed text detection and recognition framework is tested on benchmark datasets to analyse the recognition accuracy and processing time. The results show that the proposed framework is capable of achieving a higher accuracy level then the existing state-of-the-art approaches.

1.4 Organization of the Thesis

The thesis is organized into eight chapters as summarized below:

Chapter 2 presents a comprehensive review of the existing literature and state-of-the-art techniques in the research areas that have been addressed in this thesis. The chapter briefly highlights the shortcomings of the existing techniques.

Chapter 3 concentrates on providing the background knowledge related to the research context in which the novel techniques have been proposed. It covers the theoretical, conceptual and mathematical definitions utilized in chapters 4 - 7 and builds the foundation of the original contributions of this thesis.

Chapters 4 is the first contributory chapter of this thesis and presents original research in moving object detection and recognition. It describes a novel feature representation technique for recognizing humans and vehicles in the video frames. The chapter also describes novel enhancements to an existing foreground extraction technique and thus formally presents them in the context of moving object detection.

Chapter 5 presents a novel pipeline for the correction of colours in compressed CCTV videos and the work is further extended towards the accurate recognition of colour of objects present in images and video frames. The chapter presents the extensive performance analysis of the proposed pipeline on benchmark datasets.

Chapter 6 presents the original contributions in the area of vehicle make and model recognition. The research emerged into a novel visual words based textured shape representation technique. The type of vehicle is also identified based on the results of recognized model.

Chapter 7 proposes a novel methodology for the detection and recognition of text in the outdoor scenes. The chapter presents the improvement to the existing BR based character identification technique. Extensive performance estimations both in terms of accuracy and computational complexity have been presented to elaborate the effectiveness of the proposed framework.

Chapter 8 summarizes the research presented within this thesis and draws conclusions. It also presents an insight into the possible future improvements of the proposed algorithms.

Chapter 2

Literature Review

2.1 Introduction

The literature review focuses on four major research areas that the presented work in this thesis deals with:

1. Moving object detection and recognition.
2. Correction and recognition of colours in images and videos.
3. Make and model recognition of vehicles in images and videos.
4. Detection and recognition of text in outdoor scenes.

These four topics cover the contributory subjects of this thesis presented in chapters 4, 5, 6 and 7. This chapter covers the important literature regarding all the subjects in separate sections. All the research problems addressed in this thesis are of great importance to research community due to which a significant volume of literature is already published in these areas. Due to the volume of relevant publications, it is impossible to cover all the existing approaches in this literature review, however, the most popular and the most relevant of them are reviewed in this chapter.

Section 2.2 reviews the techniques related to detection and recognition of humans and vehicles. Section 2.3 gives a comprehensive review of literature related to the subject of colour correction and recognition. In section 2.4, the research work related to vehicle make and model recognition is reviewed in significant detail. The comprehensive literature related to the problem of text detection and localization is reviewed in section 2.5. Finally, section 2.6 covers the important literature related to the recognition of characters and words in the images.

2.2 Detection and Recognition of Moving Objects

The annotation process of videos strongly relies upon the accurate segmentation and classification of foreground objects. The identification process ensures that the segmented blobs are classified into humans, vehicles and background correctly. The area of identification and classification of objects has been well researched. The detection of the objects of interest is mainly carried out using one of the following techniques:

1. Sliding Window based search
2. Background subtraction

Sliding Windows: In the context of object detection, Papageorgiou et al. [109] proposed the use of sliding window based search method. The sliding window technique performs an exhaustive search for the object of interest by extracting hundreds of windows at various scales and locations of the image. Each extracted windows is classified as object or non-object on the basis of shape or texture features. These methods have been extensively used by various authors for detecting the objects of interest in the images. For instance, the research work in [163, 32, 25] used sliding windows based framework for pedestrian detection, [168, 139, 93, 94] used it for text localization, [162] performed face detection using such technique, [54] and [35] employed such a framework for vehicle localization and vehicle license plate detection respectively. The sliding window based search algorithm generates thousands of windows per image. The computation of feature vectors on each detection window and performing subsequent classification becomes a cumbersome job that requires significantly high memory and processing resources.

Background Subtraction: Background subtraction methods are widely used to detect moving objects in videos from static cameras. These methods aim at detecting the moving objects using the difference between the current frame and a reference frame often referred as background model. Several techniques for background subtraction have been proposed in the literature. The most basic technique to perform background subtraction is to use an image void of moving objects as background model [31, 57, 193]. Lo and Velastin [87] proposed to use the median value of the last n frames as the background model. Cucchiara et al. in [30] proposed to compute the median on a set of values containing the last k subsampled frames and w times the last computed median value. This combination has shown to increase the stability of the background model. Wren et al. [175]

proposed to model the background independently at each (x, y) pixel location. The model is based on ideally fitting a Gaussian Probability Density Function (PDF) on the last k pixel's values. In order to deal with complex backgrounds made of animated textures (such as waves on the water or trees shaken by wind), some authors proposed the use of multi-modal PDFs. For instance, Stauffer and Grimson [148] proposed to model every pixel with a mixture of K Gaussians. Such a model is able to cope with multiple background objects. It provides a complete description of background and foreground in an image. The proposed framework in [148] gained significant attention from research community. The Stauffer and Grimson's background modelling technique is described in detail in section 3.2. The performance and complexity of various background modelling and subtraction methods significantly varies.

In any case, the shape and/or texture features are used to represent the object of interest. The selection of feature representation technique plays a vital role in the performance of the detection framework. This section reviews the existing feature representation techniques that were developed to address the problem of humans and vehicle detection in images and videos.

2.2.1 Gradient based Techniques

The gradient based shape representation schemes gained significant attention of the research community during the last decade. The proposed techniques in [32, 170, 132] aim at characterizing the local shapes by capturing edges and gradient structures. Dalal and Triggs [32] proposed a person detection technique to detect people in an image. They developed locally normalized HOG features to capture the local gradient orientation structures to characterize human images. A Support Vector Machine (SVM) classifier was trained on a comprehensive training data and used for the classification of the extracted windows. The framework used sliding windows based strategy to compute Histogram of Oriented Gradients (HOG) features in consecutive windows of variable sizes and classified each window as a human or non-human. Beymer et al. [10] presented a method for detection of cars by locating the corner features on multiple consecutive frames. A few other methods [8, 9] also used corner features combined with edge maps in template matching scheme to detect cars in highway video scenes.

Belongie et al. [7] proposed the Shape Context (SC) features to characterize both global and local shape structures. A few other techniques [2, 19] used the geometric configuration of different local parts of objects to represent and classify those objects. These methods represent the image as a loose collection of quasi-independent local patches and compute the robust visual descriptors on these patches. A statistical aggregation process is then used to capture the statistics of the resulting set of descriptor vectors and quantification of image appearance.

Recently, Chen [25] proposed a combination of intensity-based rectangular and gradient-based 1-D features for human detection in images. The proposed method used AdaBoost algorithm for selecting the critical features from a combined feature set. A classifier is then learned using the training images for each stage of the cascaded structure. Meta-stages were introduced to enhance the detection performance of the boosted cascade.

2.2.2 Frequency based Techniques

The methods in this particular class use frequency domain to highlight and capture the pattern. These methods aim at analysing the texture present in the frequency coefficients of the image and classify the objects accordingly.

Papageorgiou and Poggio [110] proposed a trainable system for the detection of objects in unconstrained and cluttered scenes. They devised a feature descriptor using Haar features and Wavelet Transform (WT) for representing various object classes. The proposed technique use example-based learning approach where negative and positive training examples are used to learn the classification model. The method was robust to colour and texture changes. The system was tested in face, people and car detection tasks. The human detection system was trained on 1,848 positive patterns (924 frontal and rear people images and their mirror images) and 11,361 non-people patterns and tested on 123 images containing people and 794,906 non-people patterns. In the same way, the face detection system was trained on 2,429 facial images and 13,229 non-facial patterns. The test set comprised of 105 facial images and 3,909,200 non-face patterns. The car detection system was trained on a set of 1032 frontal and rear images of cars (516 examples and their mirrors) and 5,166 non-car patterns. The test set contained 90 images of cars and 600,272 non-car images. They reported a detection rate of 90% at False Positive Rate of 10^{-4}.

Viola et al. [163] integrated image intensity information with frame differencing using Haar-like wavelets and applied this method to human movement detection. They used detection style algorithm that slide a detector over two consecutive frames of video sequence. A set of 5 rectangle filters were designed to estimate the motion of objects. They employed Adaboosting to train a chain of progressively complex region rejection rules based on Haar-like wavelets and space-time difference. The proposed approach was tested on PETS 2001 dataset that contains videos of people and vehicles recorded from a camera mounted at a very high location. A detection rate of 80% at false positive rate of 1/400000 was reported.

2.2.3 Texture based Techniques

The methods in this particular domain aim at capturing the local texture in the objects for their classification. Various strategies to characterize the local texture have been proposed and applied to the problem of object detection.

A number of filters such as Gabor [33] and other linear filter-banks [173, 159, 80] along with the local descriptors such as Scale Invariant Feature Transform (SIFT) [88], colour SIFT [1], LBP [105], Speeded Up Robust Feature (SURF) [6], Maximally Stable Extremal Regions (MSER) [44], Region Covariance [154] and Spin Images [75] were used to classify the local texture and applied to the problem of object re-identification [52]. These filters or descriptors can be applied to sparse feature points or on a dense grid.

Recently, a trend has emerged to compute the texture features such as Local Binary Patterns (LBP) in the gradient map or frequency domain of images to develop specialized feature representation schemes for the detection of humans and vehicles.

Wang et al. [171] combined HOG with cell-structured LBPs as the feature set to propose a human detector with the capability of handling partial occlusion. LBP descriptors are invariant to monotonic gray level changes and they are computationally efficient. The authors designed two types of detectors; first for the whole scanning window, second for the local regions using linear SVM applied on training data. The proposed framework achieved a detection rate of 91.3% with Flase Positives Per Window (FPPW) = 10^{-6}, 94.7% with FPPW = 10^{-5}, and 97.9% with FPPW= 10^{-4} on the INRIA person dataset [32]. However, the method was computationally expensive due to high-dimensional feature space similar to [32].

To overcome the problem of high dimensional feature space, Zheng et al. [190, 191] proposed a variation of LBPs for pedestrian detection. They presented dense Centre-Symmetric Local Binary Patterns (CS-LBP) and pyramid centre-symmetric local binary/ternary patterns (CS-LTP), for pedestrian detection. The proposed feature descriptor was combined with linear SVM for classification. The experimental evaluation was performed on INRIA person dataset and reported the superiority of the method.

Kim et al. [72] proposed an approach based on the combination of a Wavelet-based CS-LBP (WCS-LBP) with a cascade of Random Forests (RF). They extracted three types of WCS-LBP features using the scanning window of WT sub-images to reduce the feature dimension. The extracted WCS-LBP descriptors were then applied to a cascade of RF. Pan et al. [108] proposed the use of LBPs in Contourlet Transform (CT) sub-bands for object classification. They analysed various statistical models generated from CT coefficients for object representation.

2.2.4 Other Techniques

There is no existing standard about the selection of descriptors to annotate humans and vehicles. Therefore, existing techniques vary in the type and number of features used to annotate humans and vehicles. Work by Hansen et al. [53] is closely related to the subject of this thesis. Their proposed system aimed at annotating the human objects with three appearance features: the primary colour of the clothing, the height of the human subject and the focus of attention. Bregler and Malik [14] proposed a technique to decompose the image into local regions using Generalized variance to identify different classes of the vehicles. The generalized second moments were derived using multiscale, multiresolution filter banks and combined with the Hierarchical Mixtures of Experts (HME) [63] to model the shapes. Lipson [84] and Lipson et al. [83] proposed a system using deformable templates for the detection of cars from side view.

Table 2.1 summarize the details of moving object recognition techniques.

Table 2.1: Summary of Object Detection techniques.

Method	Summary of the technique
Dalal and Triggs [32]	Locally normalized HOG + SVM.
Beymer et al.[10]	Detection of corner features in consecutive frames.
Belongie et al. [7]	Shape Context features.
Agarwal et al. [2], Carneiro and Lowe [19]	Geometric configuration of different local parts of objects + statistical aggregation.
Chen [25]	Intensity-based rectangular and gradient-based 1-D features + AdaBoost algorithm.
Papageorgiou and Poggio [110]	Haar features and Wavelet Transform (WT) + Exemplar SVM.
Viola et al. [163]	Haar-like wavelets and motion information.
Wang et al. [171]	Combination of HOG and cell-structured LBPs + linear SVM .
Zheng et al. [190, 191]	Dense Centre-Symmetric Local Binary Patterns (CS-LBP) and pyramid centre-symmetric local binary/ternary patterns (CS-LTP) + linear SVM.
Kim et al. [72]	Combination of a Wavelet-based CS-LBP (WCS-LBP) + cascade of Random Forests.
Pan et al. [108]	LBPs in Contourlet Transform sub-bands.

Bregler and Malik [14]	Combination of generalized Second Moments and the Hierarchical Mixtures of Experts for modelling shapes.
Lipson [84], Lipson et al. [83]	Deformable templates.

2.3 Object Colour Retrieval for Video Annotation

A substantial volume of literature is devoted to the problem of colour based content analysis, image segmentation, image retrieval, video indexing etc. According to Schettini et al. [130], the problem of colour extraction is based on either predefined conversant colours or a query illustration. They presented an extensive survey of methods for colour image indexing and retrieval in image databases. The existing techniques for colour recognition and retrieval can be roughly classified into two categories:

1. Feature space based methods.
2. Spatial domain methods.

This section reviews a number of relevant techniques belonging to both categories and analyse the advantages and disadvantages of existing methods. The aim here is to specify the potential research areas that require attention while considering the application of surveillance video annotation and retrieval.

2.3.1 Feature Space Methods

Feature space methods directly use colour histograms and their quantization to estimate the dominant colour of a particular region within an image. These methods are mostly used in image indexing.

Swain and Ballard [150] proposed the idea of colour recognition for indexing applications based on colour histograms that were matched by a technique called histogram intersection. They demonstrated that colour histograms of multi-coloured objects provide a robust and efficient cue for indexing. The histogram intersection technique proposed in their work matched model and image histograms in fast incremental order that made the system suitable for real-time video indexing in a large database of stored models.

Saber et al. [121] proposed an adaptive colour classification technique for automatic image annotation. The technique used YUV colour space and modelled the colour distribution within each colour class using Gaussian Probability Density Function (PDF), the mean and covariance matrices were computed using an appropriate training set. The model defined a mapping criteria from 2-D chrominance space to a scalar γ. Binary hypothesis test was applied in succession to decide whether a particular pixel belong to any of these classes. The adaptive thresholds for different classes were calculated as the functions of the histograms of the γ and the universal thresholds, that were computed from Receiver Operating Characteristics (ROC) curves of training images. Finally the image regions were assigned to the keywords to annotate the content based on their colour.

Smith and Chang [145] proposed a simple method for automatic colour extraction and indexing. The proposed system identify regions in the image containing colours using predefined colour set. They quantized the Hue-Saturation-Value (HSV) space into 162 bins (18 Hue, 3 Saturation and 3 Value) and assigned pixels to their respective bins after applying a Median Filter (MF) to reduce the impact of small detail and spot noises in the image. If a region contained one of the colours from predefined colour set and the colour pixel constituted more than 20% of that region area then that colour was assigned as the representative colour of that region. The paper did not presented any quantitative results.

Son et al. [146] proposed a convolution Grid Kernel (GK) based method for vehicle colour recognition. Their proposed system transformed input image to HSV space and generate colour recognition models for predefined colour classes of training images using SVMs with Grid Kernel. The GK compares the pixels using H and S components to reduce the effect from surrounding noise. It mapped images onto a high dimensional feature space of which features were image fragments. The similarity between two images (training image and test image) was obtained by the inner product of two image vectors. The GK based approach achieved 92% precision and 92% recall on the dataset of 500 vehicle images.

Payne et al. [111] represented every pixel with a feature vector containing H, S and V components along with the location (x, y) information. The colour clustering was performed using lossy compression method based vector quantization technique that aimed at minimizing the total sum of squared quantization errors. The proposed vector quantization was performed in multi-scale arrangement of images that was achieved using Gaussian pyramid. The vector quantization algorithm determined the appropriate centroids to assign to all pixels in the image. The colour space and the spatial information in the feature vector helped to ensure that each image region get homogeneous colour distribution after performing the colour clustering. Lastly, a tracing algorithm ensured that each extracted region comprise of contiguous set of pixels.

The feature space techniques are simple and do not require high resource utilization. However, the choices related to number of quantization levels and the selection of colour space is different and no standard exists to ensure the performance of colour categorization in all applications.

2.3.2 Spatial Domain Methods

Spatial domain techniques are mostly used for colour based image segmentation tasks. These techniques work on every pixel and perform region growing in an iterative manner to form significant regions of pixels with nearly the same colour.

Ilea and Whelan [62] proposed a spatial k-means clustering based technique to segment image regions using colours. They devised an adaptive filter that took into account not only the colour but also the texture of the region to compute the dominant colour. They also proposed diffusion based filtering for image smoothing in order to reduce the noise. The initial centres for the k-mean clustering were computed using the peaks in the histograms of R, G and B components separately. The method was tested on different colour spaces; however, the paper only presented visual results to show the performance of the proposed technique.

In a study in [74], the authors presented solution to various problems encountered in real-time colour recognition of objects. They showed that the use of thumbnails instead of actual sized frames not only reduce the processing requirements but the overall colour recognition efficiency remain unaffected. The study emphasized the use of linear colour space (such as HSV, YUV etc.) instead of non-linear colour space (such as RGB). Lastly, the study showed that the use of Hard C-Mean (HCM) has an advantage over Fuzzy C-Mean (FCM) in terms of accuracy to find dominant clusters in the data of different colour pixels. The proposed bits were applied on a small collection of surveillance videos to recognize the colour of passing vehicles. A recognition accuracy of 90% was reported.

Weijer et al. [156] proposed a framework to learn colour names from real world images. They presented an adapted version of Probabilistic Latent Semantic Analysis(PLSA) [58] to discover colours in a bag of pixels representation, where every pixel was represented by its Lab space value. A training set was used to learn colour names and model every image as being generated by two distributions: the foreground distribution which was determined by its colour name label and the background distribution which was shared between all images. Two separate ways were adopted to assign colour names to individual pixels; the first one used the pixel value and the second one used the pixel value along with its surrounding region. A recognition accuracy of 82% was reported on a dataset of 440 human labelled images collected from eBay.

Zhang et al. [189] proposed PLSA model based approach for object colour categorization in videos taking vehicles as objects of interest. The proposed framework computed Scale Invariant Feature Transform (SIFT) features on MSERs to articulate the vehicles into various parts. A spatial shape PLSA model was then used to categorize the object parts into foreground (areas that contribute to the colour of the object) and background (area that do not contribute to the main colour i.e. tyres, windows, wind screen and background noise etc.), whereas, a spatial colour PLSA model was used to predict colour. Using joint PLSA model, the probability of foreground for each pixel in the image was computed by averaging the foreground probability of the regions covering that pixel. In this way a background subtraction was performed and HSV histograms for estimated foreground regions were constructed to compute the colour of that region by quantizing the HSV space into 8×8 bins. A dataset containing 435 car images collected from outdoor videos was used for experiments and an average colour classification accuracy of 95.8% was reported.

As expansion to their work in [188], the authors replaced the use of discriminative classifier (i.e. SVM) to improve the accuracy of colour categorization of foreground objects in surveillance videos. They proposed a two stage classification scheme; the first stage classified the regions into the colourless and colourful categories and the second stage classified the region into seven colour categories (i.e. black, white, grey, red, yellow, green and blue). They also incorporated weather conditions (i.e. sunny, cloudy or rainy) classification step to improve the colour extraction results. The experiments were performed on a dataset of 15456 images collected from surveillance videos recorded under various weather conditions. Out of 15456, 12011 frames were used for training and 3445 were used for testing. An average classification accuracy of 80% was reported.

2.3.3 Other Techniques

This section discuss the methods that are either formed by fusing the methodologies of both categories explained above or use a completely different approach for performing the colour retrieval in videos and images.

Chen et al. [23] combined feature space information and spatial domain information for a fast image segmentation method. In their work, they transformed the image into HSV space and applied k-means clustering on quantized HSV histogram to reduce the computational complexity. The initial centres and cluster numbers were dynamically computed using maximin algorithm. Each colour histogram bin was classified to its nearest cluster centroid on the basis of Euclidean

Distance (ED). Image pixels were labelled with the index of the nearest centroid of their corresponding histogram bins in an iterative manner until the process converged. Finally a statistical filter was applied to eliminate noise and unnecessary details of the labelled images. The smaller spatial regions were merged into bigger neighbours to avoid over-segmentation.

Li et al. [81] used the direct matching of histograms of vehicle images (in various colour spaces) with the training images of various colours. They used minimum distance measure (i.e. ED) to compute the similarity between histograms to decide vehicle colour. In their experiments, which were performed on vehicle images, they showed that HSI and CIELab colour space had higher recognition rates.

Brown [15] proposed a retrieval system for extracting the coloured objects from surveillance videos. The method included two sets of parameters; the first set was used to compensate for illumination conditions and camera differences and the second set for tuning the colour extraction for specific object types and optimal retrieval. The framework accumulated HSI histogram of objects over several frames. The dominant colour of each object was extracted using a proposed set of rules to quantize the HSI space. The quantization procedure was devised using three dominant colours and their proportions in the histogram. The proposed methodology was tested on videos of on road vehicles and a correct classification rate of 80% was reported.

Wui et al. [177] addressed the task of colour classification into pre-specified colours for tracked objects. They analysed significant number of video clips collected under various lighting conditions and distances from several video cameras. They computed the drift in the colour space for eleven significant colours in the surface of moving objects. The drift patterns for each representative colour were learned for the classification of unseen surface colours. Finally, a distance function was devised to perform colour identification and matching. Colour recognition accuracy of over 95% was reported on a collection of real surveillance videos.

A number of well-known approaches have been discussed in this section. Table 2.2 summarize the details of colour recognition methods.

Table 2.2: Summary of Colour Recognition Methods.

Method	Color Spaces	Summary of the technique
Swain and Ballard [150]	RGB	Histogram Intersection and Back projection for colour classification. The histograms used were created from the combinations of RG, BRG and RGB.

Saber et al. [121]	YES	2-D Chrominance information is mapped to a scalar and each pixel is matched with already learned Gaussian PDFs of training image classes using binary hypothesis.
Smith and Chang [145]	HSV	Quantized HSV space into 166 bins. The regions were indexed on the basis of size and the amount of contribution of various colours within that region.
Son et al. [146]	HSV	An SVM model is learned for HS space of training colour classes using Grid Kernel. Classification of various colours images is performed using the learned SVM model.
Payne et al. [111]	HSV	Pixels are represented by a vector that contains H, S and V values along with location information. Colour clustering is performed by quantizing these vectors into clusters using a similar approach as K-mean clustering.
Ilea and Whelan [62]	RGB, YIQ, HSI	K-mean clustering is used to segment regions of image with similar colours. Initial cluster centres are selected using R, G and B histograms.
Kuo et al. [74]	HSV, YUV	Hard C-mean clustering to find the cluster centres of dominant colours.
Zhang et al. [188, 189]	HSV	A joint PLSA (comprised of Shape PLSA and Colour PLSA) is used to classify each pixel as foreground or background.
Weijer et al. [156]	Lab	PLSA analysis is performed to determine if a pixel is a background or foreground, where every pixel is represented by its Lab values.
Chen et al. [23]	HSV	HSV space is quantized and K-mean clustering is applied to the clusters of bins.
Li et al. [81]	Various	Histograms of test and training images are matched using Euclidean distance measure to decide about the colour of the test image.

Brown [15]	HSI	Histogram over various frames is computed and three dominant colours and their compositions are found. A set of devised rules is applied to finalize the colour of an object.
Wui et al. [177]	HSL	Drift patters for various colour pixels are computed in videos and an SVM is trained for each training colour. Test colour sample is classified using the SVM score.

2.4 Vehicle Make and Model Recognition

The problem of Vehicle Make and Model Recognition (VMMR) is a well addressed problem in literature, however, existing techniques were developed for specific scenarios. Most of the existing techniques deal with the recognition of vehicles in videos. A methodology that may be considered as a standard baseline solution to handle most practical scenarios is yet to be emerged. The existing approaches can be broadly categorized into three categories:

1. Shape based techniques.
2. Frequency based techniques.
3. 3D Modelling and analysis based techniques.

The well known approaches from each category have been discussed in the following sub-sections.

2.4.1 Shape and Texture based Techniques

Pearce and Pears [112] proposed a technique that recursively partitioned the image into quadrants; the strength of frequency features in these quadrants were then summed and locally normalised in a recursive hierarchical fashion. They also tested two classification techniques; a k-Nearest-Neighbour (kNN) and a Naive Bayes classifier. An accuracy of 96% was reported using leave one out cross validation on a dataset of 262 frontal images of cars.

Petrovic and Cootes [115] investigated a number of features including direct and statistical mapping methods to extract the information from rigid structures such as headlights and grill in car images. The recognition process relied on a reference segment (the front number plate). The location and scale of the reference segment was used to extract Region of Interest (RoI) in the image from which the structure was sampled. Feature vectors were classified using kNN classification. A recognition rate of 93% and verification equal error rate of 5.6% was reported on a collection of more than 1000 images containing 77 different classes. The system robustly tackled a wide range of weather and lighting conditions, however, no statistical measures were reported.

Munroe and Madden [97] investigated three classification strategies; kNN, feed forward neural network and a decision tree using thickened Canny edges as feature descriptors for vehicle images. An identification rate of 97.46% was reported with the help of kNN classifier on a test set of 150 vehicle images.

Clady et al. [27] proposed the use of oriented contour pixels to represent vehicle classes. For each class an array was formed using the oriented contour points that were stable across the class training samples. These contour points were then used to vote on whether or not a sample belongs to that class. Like [115], their system also relied on the license plate for the selection of RoI. A set of 291 high quality frontal vehicle images distributed in 50 classes were used for training. An identification rate of 93.1% was reported on a test set of 830 outdoor vehicle images with variations in lighting, angle, distance and resolution.

Lee et al. [79] proposed an image rectification method to improve RoI extraction by reducing the effect of skew. The initial RoI was extracted using licence plate as a reference segment. They further used HOG as visual descriptor and tested kNN and SVM for classification of test images.

Psyllos et al. [118] proposed a hierarchical method where make recognition and model recognition were treated as separate problems. The technique extracted RoI from the vehicle's front view using the location of the license plate as a reference segment. In the next step, manufacturer's logo was extracted with reference to the location of RoI. Probabilistic Neural Network (PNN) in conjunction with phase congruency features were computed to recognize the vehicle's make (e.g. Ford, Renault etc.). In order to determine the vehicle's model, SIFT features were used to represent RoI of vehicle image and classified using a PNN. A dataset comprised of 110 (55 training and 55 test) frontal view vehicle images was used to test the performance of the system. Recognition rate of 85% for make and 54% for model was reported along with 90% colour recognition accuracy.

Daya et al. [34] proposed the use of geometric parameters to represent the different makes and models of vehicles. In order to represent the vehicles, three geometric measures were computed from the RoI of vehicle's frontal view. These geometric measures were normalized using the height of the licence plate and classified using a Neural Network. A recognition rate of 95% was reported on a dataset containing 12 test classes.

Sarfraz et al. [127] proposed the use of Local Energy Based Shape Histogram (LESH) features to represent different vehicle classes. LESH features were computed on RoIs taken from the front view of the vehicles. The extracted features were modelled in a similarity feature space using a probabilistic Bayesian framework. Using Bayes rule, the posterior over possible matches was computed and the highest score was selected as the make and model class. High recognition rate of 94% was obtained in experiments. This approach offered the advantage that only a single image per class was required as a reference after offline training.

Sarfraz et al. [126] further proposed the use of salient regions called "Patches" to represent vehicles and non-vehicle regions. A local description for each patch was extracted using LESH features. The local description was modelled in a similarity feature space classified using a probabilistic Bayesian framework. An accuracy of 94% on the dataset of [118] and 62% on another dataset that was collected in uncontrolled conditions was reported.

Saravi and Edirisinghe [123] presented the use of temporal information to improve recognition accuracy of vehicles in CCTV videos. They used Coherent Point Drift (CPD) to remove skew in the image using the four corners of license plate as reference. They employed LESH features for vehicle images representation and classification procedure was performed using trained SVM. A recognition accuracy of 94% on a dataset of images and 66.6% on a video dataset was reported.

In recently published approaches, Chen et al. [22] proposed a dynamic sparse representation scheme to represent a vehicle model in an over-complete dictionary using the training vehicle images. Hamming distance classification scheme was used to classify vehicle's make and model to detailed classes. They exploited the vertical symmetry of pattern on two halves of the front side of vehicles and devised a symmetrical Speeded Up Robust Features (SURF) descriptor. All possible matching pairs of SURF points were used to detect vehicles in the image and extract RoI. A collection of 2440 training images and 3738 test images was used to test the performance of the system. Vehicle detection accuracy of up to 98.85% and recognition accuracy of 94.60% was reported on various car models under varying illumination conditions.

2.4.2 Frequency based Approaches

Kazemi et al. [70] presented a comparative study of the performance of Fourier, Wavelet, and Curvelet transform features for VMMR. In experiment, the transforms were applied on a RoI and the coefficients of the transforms mentioned above were fed into a kNN classifier. The best recognition rates were achieved using all the Curvelet Transform coefficients as features. They examined only 5 vehicle classes. As an extension of their work in [71], the authors proposed the use of trained SVM on rear view images of the vehicles. They used Curvelet Transform based features to represent various classes of vehicles. The standard deviations of the curvelet coefficients for each scale and direction were used as features. The experiments were done with a kNN classifier and two SVM classifiers in one-versus-one and one-versus-all manner. A dataset comprised of 230 training images and 70 test images was used for experiments and a detection accuracy of 99% was reported.

Rahati et al. [119] proposed a similar technique as [70]; they used CT sub-bands instead of Curvelet Transform. The CT coefficients at different sub-bands and directions were computed. The standard deviations of CT coefficients from selected scales with the most useful information were fed to the SVM classifier. A recognition rate of 99% was reported on the dataset presented in [70].

Zafar et al. [186] proposed an improvement to the technique presented in [119] by making use of localized CT feature extraction instead of standard deviations of the CT coefficients. A trained SVM model was used for the classification of vehicle models. A dataset containing 250 training images and 50 test images was used for experimentation and a recognition rate of 94% was reported. An accuracy of 52% was reported when the technique in [119] was tested on the mentioned dataset.

2.4.3 3D Modelling based Technique

Recently, the paradigm has shifted towards 3D modelling and analysis for object detection, segmentation and recognition. A few approaches [117, 120] that incorporate 3D modelling and analysis have been proposed to address the problem of VMMR.

Prokaj et al. [117] presented the use of prior knowledge of the ground plane to impose constraints on the virtual camera motion. The vehicle pose from each frame of the video was estimated and its 3D motion on the plane was calculated using a structure from motion algorithm. The 3D models of vehicles in database were rotated to the same pose as the calculated 3D structure; then features similar to SIFT features were extracted but with rotation invariance disabled. Using these features, a video-model similarity metric was computed and the model with the highest score was selected as the make and model of the test image. The system was tested on a dataset of 20 video clips containing 36 vehicle classes and a recognition rate of 50% was achieved.

Ramnath et al. [120] presented the 3D curve alignment based technique for VMMR. They used 3D space curves obtained by back-projecting image curves onto silhouette based visual hulls and then refined them using three-view curve matching. The acquired 3D curves were then matched to 2D image curves using a 3D view-based alignment technique.

Recently, a number of texture based object recognition and image classification approaches [142, 122, 104, 76] have been proposed that use densely computed features. These features are extracted all over the image and encoded into a more meaningful representation.

The densely computed features are usually encoded using Mid-Level Representation (MLR) techniques that summarize the dense description into single feature vector that is suitable for statistical learning and recognition. Bag of Visual Words (BoVW) model [29] is the most widely used encoding scheme that computes the occurrences of vector-quantized descriptors, however, in recent literature various new encoding schemes have been emerged that aim at reducing the loss of information in the vector quantization step in BoVW. These include Locality Constrained Linear (LLC) encoding [167], Vectors of Locally Aggregated Descriptors (VLAD) [67], Super Vector Coding [194] and Fisher Vectors (FVs) [64, 114]. Chatfield et al. [20] evaluated the recent feature encoding methods on a number of image recognition benchmark datasets and showed that the performance of FV has out-performed all other feature encoding techniques.

Table 2.3 summarizes the details of VMMR techniques.

Table 2.3: Summary of Vehicle Make and Model Recognition techniques.

Method	Summary of the technique (Image Representation + Classification)
Petrovic and Cootes [115]	Feature descriptors from rigid structures such as headlights and grill in car images + kNN.
Munroe and Madden [97]	Thickened canny edges + kNN.
Clady et al. [27]	Oriented contour pixels + voting strategy to classify query images.
Psyllos et al. [118]	Phase congruency features and SIFT features + Probabilistic Neural Network.
Sarfraz et al. [127]	LESH + Probabilistic Bayesian framework.
Saravi and Edirisinghe [123]	LESH + SVM.
Chen et al. [22]	Dynamic sparse representation + Hamming distance.
Kazemi et al. [70]	Curvelet Transform + kNN classification.
Rahati et al. [119]	Contourlet sub-bands + SVM.

Prokaj et al. [117]	Structure for Motion for pose estimation + SIFT descriptors for classification.

2.5 Text Detection and Localization

A significant volume of literature concentrates on detection of text from images and videos. Existing text detection approaches can be classified into three broad categories:

1. Texture based approaches.
2. Connected Component (CC) based approaches.
3. Object detection based methods.

Few well-known methods in these categories are briefly discussed in the following sub-sections.

2.5.1 Texture based Techniques

In texture based techniques, Jain and Zhong [65] used the distinguishing texture present in text to determine and separate text, graphics, and half tone image regions in scanned grayscale document images. Further in [192], they utilized the texture characteristics of text lines to extract text in grayscale images with complex backgrounds. In contrast to that, Wu et al. [179, 178] segmented input images using a multi-scale texture segmentation scheme. Potential text regions were detected based on nine second-order Gaussian derivatives. Sin et al. [143] used frequency features such as the number of edge pixels in horizontal and vertical directions and Fourier spectrum to detect text regions in real scene images.

Mao et al. [90] proposed a texture-based text localization method using WT. Lim et al. [82] made a simple assumption that text usually has a higher intensity than the background. They counted the number of pixels that were lighter than a pre-defined threshold value and exhibit a significant colour difference relative to their neighbourhood. A candidate region with a large number of such pixels was regarded as a text region. Lee et al. [78] proposed the use of SVM and spatio-temporal restoration for text detection in videos. Chen and Yuille [24] employed the AdaBoost algorithm and the joint probabilities of the features including X and Y derivatives, histogram of intensity and edge linking to detect the text in natural scenes for visually impaired persons.

Ye et al. [182] employed multi-scale wavelet features to locate text line in the presence of complex background. The advantage of texture based detection methods is that they perform well in the presence of noise and cope with illumination inconsistencies in the image. However, the big disadvantage of texture based text detection methods is their computational complexity. A few authors proposed various strategies to speed up the process by selecting RoIs, yet, the methods in this category are computationally expensive than CC based methods. This disadvantage makes them less suitable for practical systems and real time applications.

2.5.2 Connected Component based Methods

In CC based techniques, Ezaki et al. [41] combined edge image, reverse edge image and colour based analysis for CCs extraction. The top scoring contestant in ICDAR05 [89] challenge applied an adaptive binarization method to find CCs. Shivakumara et al. [140] used k-means clustering in the Fourier-Laplacian domain for CC extraction. Epshtein et al. [40] proposed a novel image operator named as Stroke Width Transform (SWT) that seek for the value of stroke width for each image pixel. They demonstrated its use for the problem of text detection in natural scene images. The proposed operator showed promising results and acquired significant attention from research community mainly because of its simplicity and robustness to detect text in many fonts and languages. However, the SWT method relies on Canny edge detection and its performance deteriorate in images containing high noise and illumination variation.

Yao et al. [180] proposed a two level classification scheme and two sets of features (component level and chain level) for capturing both intrinsic characteristics of text regions. The proposed framework used SWT for finding the potential candidate regions in the images. Mosleh et al. [96] proposed a novel bandlet-based edge detector to enhance the accuracy of SWT that originally uses Canny edge detector. Neumann and Matas [99, 101] used MSERs to detect potential candidate regions. Chen et al. [21] used edge enhanced MSERs to find letter candidates. In recently proposed methods, Neuman and Matas [100] used various properties of strokes to localize and recognize characters. Huang et al. [61] proposed stroke feature transform filter that extends the capability of SWT by incorporating colour cues of the pixels.

A considerable advantage of CC based techniques is their computational performance as compared to texture based techniques. On the other hand, these techniques perform poorly in the presence of noise because they mainly rely on gradient and edge images that are sensitive to minor changes.

2.5.3 Object Detection based Methods

Apart from texture and CC based methods, a new direction, that applies object detection techniques to locate text in the images, has recently emerged. The text is considered as an object and is located in an object detection framework. Wang et al. [168] proposed the use of famous multi-scale sliding window technique to localize text regions in the image. The technique is adopted by a number of recently proposed frameworks [139, 93, 94] for text localization.

Table 2.4 summarizes the various techniques for text detection in images.

Table 2.4: Summary of techniques.

Method	Summary of the technique
Jain and Zhong [65], [192]	Exploitation of texture characteristics of text lines for it's detection.
Wu et al. [179, 178]	Multi-scale texture segmentation using nine second-order Gaussian derivatives.
Sin et al. [143]	Frequency features and Fourier spectrum to detect text regions.
Mao et al. [90] and Ye et al. [182]	Multi-scale Wavelet Transform.
Lim et al. [82]	Image thresholding using intensity and colour information.
Lee et al. [78]	spatio-temporal restoration + SVM.
Chen and Yuille [24]	Joint probabilities of X and Y derivatives, histogram of intensity and edge linking + AdaBoosting.
Ezaki et al. [41]	Edge image, reverse edge image and colour based analysis for CCs extraction.
[89]	Adaptive binarization for candidate extraction.
Shivakumara et al. [140]	k-means clustering in the Fourier-Laplacian domain for text candidate extraction.
Epshtein et al. [40]	Stroke Width Transform (SWT).
Yao et al. [180]	SWT + Component level and Chain level analysis for text detection.
Mosleh et al. [96]	Bandlet-based edge detector based SWT.
Neumann and Matas [99, 101, 100] and Chen et al. [21]	MSERs + SVM.
Huang et al. [61]	Colour cue based stroke feature transform.

Wang et al. [168] and [139, 93, 94]	Sliding window based search.

2.6 Character and Word Recognition

A significant volume of literature exists that deals with the problem of character and word recognition in natural scene images. A number of specialized feature representations, binarization techniques, segmentation methods and word models have been proposed to date, yet, the problem of text recognition is open. The reason being the diversified nature of text and the presence of high inter-class similarity as well as high intra-class variation. The following paragraphs briefly cover the most recent work in character and word recognition.

2.6.1 Character Recognition

In the area of character recognition, Campos et al. [18] introduced Chars74k dataset of characters collected from natural scene images. They showed that commercial OCR engines do not achieve good accuracy for natural scene character images, therefore they proposed a Multiple Kernel Learning based method. Wang et al. [169] proposed the use of HOG features together with Nearest Neighbour classifier and showed the improved performance. They enhanced their work in [168] where they used Bayesian inference and show a considerable performance improvement on ICDAR03-CH dataset.

Sheshdari et al. [137] used HOG feature and exemplar SVMs and affine warping to demonstrate improved performance. Yi et al. [185] presented a comparative study about the performance of local and global HOG features for character recognition. A few word recognition and end-to-end scene text recognition methods [139, 77] reported character recognition scores separately. The most recent work in character recognition is presented by Lee et.al [77]. They used discriminative region based feature pooling to learn the most informative sub-regions of each character within a multi-class classification framework. They reported the state-of-the-art recognition performance for scene characters.

2.6.2 Word Recognition

In the area of word recognition, a number of approaches have emerged that focus on specialized modules for word recognition. For instance, Smith et al. [144] proposed a similarity expert algorithm to remove the logical inconsistencies in an equivalence graph and perform search for the maximum likelihood interpretation of a sign as an integer programming. The work in [139, 94, 93] build Conditional Random Field (CRF) models on the potential character locations in a sliding window search and add the linguistic knowledge and spatial constraints to compute pairwise costs for word recognition. The work in [168, 169] used pictorial structures to detect words in the image. The pictorial structures found an optimal configuration of a particular word using the scores and locations of the detected characters.

Weinmann et al. [172] formulated Markovian model score on the segmented candidates on the basis of appearance, geometric and linguistic characteristics. Neuman et al. [101] proposed a text recognition system where they extract potential candidate characters using a set of Extremal Regions (ER) and then perform exhaustive search with feedback loops to group ERs into words and recognize them in an OCR stage that was trained by using synthetic fonts.

Yao et al. [181] proposed a new feature representation technique named as Strokelets that captures the essential sub-structures of characters at different granularities. An important challenge in word recognition frameworks is the identification of characters in cropped word images. A number of approaches have been proposed so far to deal with this challenge. The proposed methods involve CCs [180], image binarization [172, 91], Extremal Regions [101], Graph- Cuts [92], Sliding Windows [168] and k-means clustering [164].

Recently, Field et al. [42] proposed BR for character identification. It uses colour clustering as a starting point to fit a regression model for each image and separate foreground pixels from background using an error threshold. The method reports a superior character segmentation performance in comparison to other existing techniques.

2.7 Summary and Conclusion

A brief review of the state-of-art techniques and most widely used methods for object identification, colour recognition, vehicle make and model recognition and text detection and recognition presented in this chapter. From the literature review, as documented it is obvious that there are several research gaps that can benefit from further research.

In object recognition, the proposed techniques in literature exploit gradient, texture or frequency properties to enhance and capture the shape patterns of the objects such as humans and vehicles. In recently proposed techniques the texture in the WT coefficients have been used to develop feature representations. However, the WT lacks the ability to capture the texture around the contour of the objects that possess curves in their silhouette (e.g. humans). In WT filter bank, the filters are designed to capture only those pixels (such as line segments) that strictly possess the orientation that the filter is trained to capture. The CT on the other hand provides the advantage that it captures the contours instead of pixels or straight lines. Therefore, it can be exploited to capture the silhouette of the objects that possess curves in their contours.

In colour recognition, the existing approaches involve pixel clustering based on histograms or feature space analysis to analyse and combine the regions of similar colours. However, a research gap exists especially in the case of medium to low quality surveillance videos where most of the content is dominated by noise. In such cases, a colour correction mechanism can play an important role in improving the colour recognition accuracy.

In the area of VMMR, all the existing approaches have been applied under controlled conditions and does not offer robustness to angle changes and noise. A research gap exists for a robust technique that can offer high recognition accuracy in the presence of aforementioned challenges. In addition to that there is a need for up to date benchmark image and video datasets to aid the research in this area.

In the area of text detection and recognition, none of the existing techniques have exploited the colour information up to full extent to solve this problem. Mostly the recent techniques focus on complex object detection and recognition based strategies. These techniques are accurate but at the same time require significantly high hardware resources and computation time which make them less attractive for video analysis applications.

Considering the shortcomings of the existing techniques, a number of novel frameworks for moving object detection and recognition, colour correction and recognition, vehicle make and model recognition and text detection and recognition have been proposed in chapters 4-7. The research motivation behind each approach has been presented in the relevant chapter. Chapter 3 introduces the reader to the fundamental concepts and background preliminaries that are used in the contributory chapters mentioned above.

Chapter 3

Theoretical Background

3.1 Introduction

This chapter covers the theoretical background and important concepts that have been used in the development of the proposed techniques in contributory chapters. It starts with an explanation of Contourlet Transform (CT), which is used for the classification of moving objects in to humans and non-humans. The CT have also been used in conjunction with the LBP in object identification. Consequently a section is dedicated to provide the reader with a theoretical understanding of the LBP. Next, the brief introduction of various colour spaces is given along with colour constancy algorithm and MPEG-7 Dominant Colour Descriptor (DCD). The theoretical preliminaries about the Bilateral Filter (BF) are also included in this chapter. A brief mathematical and conceptual detail about the Sacle Invariant Feature Transform (SIFT), Histogram of Oriented Gradients (HOG) and Local Energy based Shape Histogram (LESH) features is presented as these techniques have been used in the contributory work of Vehicle Make and Model Recognition (VMMR) and text detection and recognition. Lastly, the reader is introduced to the concepts and mathematical background of Fisher Vector (FV) encoding which is an important concept due to its primary importance in the computation of visual words in vehicle make and model recognition work.

3.2 Background Modelling using Mixture of Gaussian Distributions

C. Stauffer and W. Grimson in [148, 149] presented a revolutionary technique for modelling the background using Mixture of Gaussian (MoG) distributions. The technique gained significant attention from research community because of its stable performance under various scenarios such as dynamic background and gradual illumination changes. The technique models each pixel in the frame by weighted MoG distributions. The parameters and weights of these distributions are updated with each incoming frame. The updated probability of each pixel with each incoming frame is used to classify the pixel as background or foreground.

The history $\{H_1, ..., H_t\} = \{F(x_0, y_0, i) : 1 \leq i \leq t\}$ of the values of a pixel of frames F is modelled by a mixture of K Gaussian distributions. H_t can be a scalar if the pixel has one grayscale intensity value or a vector if the pixel has three colour components. The probability of observing H_t at time t is:

$$P(H_t) = \sum_{k=1}^{K} \gamma_{k,t} * \psi(H_t, \mu_{k,t}, \Sigma_{k,t}) \tag{3.1}$$

where $\gamma(k, t)$ are the coefficients of each distribution that act as weights, $\mu(k, t)$ and $\Sigma(k, t)$ are the mean value and the covariance matrix of the kth Gaussian distribution. The Gaussian function ψ is defined as:

$$\psi(H_t, \mu_t, \Sigma_t) = \frac{1}{(2\pi)^{n/2} |\Sigma_{k,t}|^{1/2}} e^{-\frac{1}{2}(H_t - \mu_t)^T \sum_t^{-1} (H_t - \mu_t)} \tag{3.2}$$

The algorithm assumes that R,G and B components are independent and have same variance values. Thus the covariance matrix is:

$$\sum\nolimits_{k,t} \left(\sigma_k^2 F\right) \tag{3.3}$$

Each new H_t is checked against the existing K distributions via absolute distance to find a match using the following equation:

$$|H_t - \mu_{k,t-1}| \leq 2.5\sigma_{k,t-1} \tag{3.4}$$

If the condition in the above equation hold then the kth Gaussian is updated using the following equations:

$$\begin{aligned} &\mu_{k,t} = (1 - \rho)\, \mu_{k,t-1} + \rho H_t \\ &\sigma_{k,t}^2 = (1 - \rho)\, \sigma_{k,t-1}^2 + \rho(H_t - \mu_{k,t})^T (H_t - \mu_{k,t}) \\ &\psi_{k,t} = (1 - \rho)\, \psi_{k,t-1} + \alpha(\beta_{k,t}) \end{aligned} \tag{3.5}$$

where $\rho = \alpha(\psi H_t | \mu_k \sigma_k)$ is a learning factor, α is the learning rate and $\beta_{k,t}$ is 1 for the model that found as matched and 0 for the remaining models. In case where no matching distribution is found, the pixel is labelled as foreground and the least probable distribution is substituted by a new one, initialized with prior parameters for $\gamma_{k,t}$ and $\sigma_{k,t}^2$ and $\mu_{k,t} = H_t$.

The distributions are arranged in an order according to the ratio γ/σ in order to decide that which portion of the distributions represents the background and which the foreground. The first ζ distributions, estimated using the following equation, account for the background:

$$\varsigma = \arg\min_{f} \left(\sum_{k=1}^{f} \psi_{k,t} > T \right) \tag{3.6}$$

where T is the threshold that defines the portion of distributions that are accounted for by the background. A pixel is labelled as background if the current H_t matches any of the first ζ distributions. In the opposite case the pixel is labelled as foreground.

This algorithm is used in chapter 4 along with proposed modifications according to the requirements of the proposed frameworks.

3.3 Contourlet Transform (CT)

Contourlet Transform was proposed by Do and Vetterli [38] in its discrete form as a simple directional extension of Wavelets. It offers a high degree of directionality and anisotropy. It provides improvements to 2-D separable Wavelet Transform for representing images with smooth contours in all directions (see figure 3.1). The idea of the CT is attributed to the grouping of nearby wavelet coefficients as they are locally correlated due to the smoothness of the contours. The CT explained by Pan et al. [108] allows for different and flexible number of directions at each scale. The image is decomposed into several directional sub-bands at multiple scales in the frequency domain by first applying a Laplacian Pyramidal (LP) multi-scale decomposition to capture the point discontinuities. A critically sampled directional filter bank is then applied on each un-decimated high-frequency band to link point discontinuities into linear structures. The overall result is an image expansion using the basic elements such as contour segments, and thus are named contourlets.

3.3.1 Pyramidal Frames

The first step as mentioned above is the multi-scale decomposition of the image to capture the point discontinuities. The authors in [38] critically analysed the LP decomposition introduced by Burt and Adelson [17] and found a drawback of the method that is the implicit oversampling. They further studied the LP in [37] using the theory of frames and oversampled filter banks. They showed that LP with orthogonal filters provide a tight frame with frame bounds that are equal to 1 and hence proposed the use of the optimal linear reconstruction using the dual frame operator (or pseudo inverse) as shown in figure 3.2.

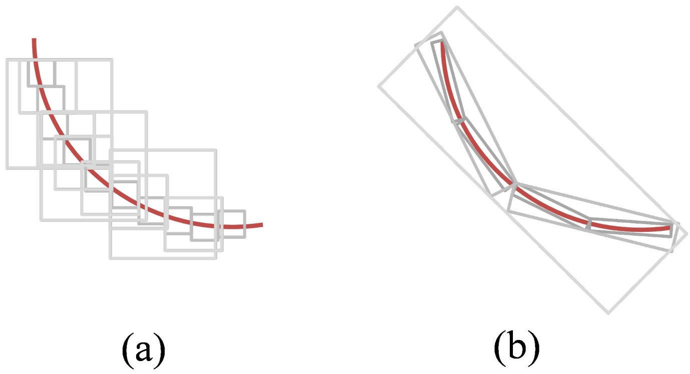


Figure 3.1: Illustration of successive refinement by (a) Wavelet, and (b) Contourlet, near a smooth contour.

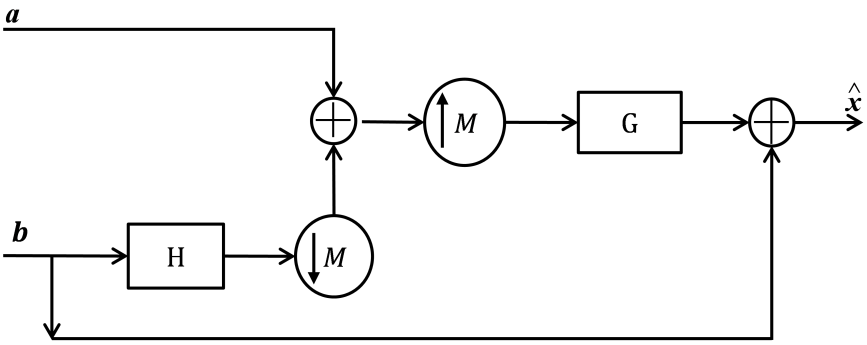


Figure 3.2: Laplacian Pyramid proposed in [37], here H and G are low pass analysis and synthesis filters and M is the integer matrix.

3.3.2 Iterative Directional Filter Banks

The Directional Filter Bank (DFB) is important to link point discontinuities into linear structures. Bamberger and Smith [4] proposed a 2-D directional filter bank by efficient implementation of an l-level binary tree decomposition, with $2l$ subbands with wedge-shaped frequency partitioning. The construction of DFB involves modulating the input image and using quincunx filter banks with diamond-shaped filters. In [36], a new implementation of the DFB is proposed that avoid modulating the input image. The new implementation uses a two-channel quincunx filter bank [161] with fan filters to divide a 2-D spectrum into two directions. Next a shearing operator is applied for reordering of the image samples. The aim for DFB here is to use an appropriate combination of the shearing operators along with two-direction partition of quincunx filter banks at each node in a binary tree-structured filter bank to obtain the desired 2-D spectrum division as shown in figure 3.3.

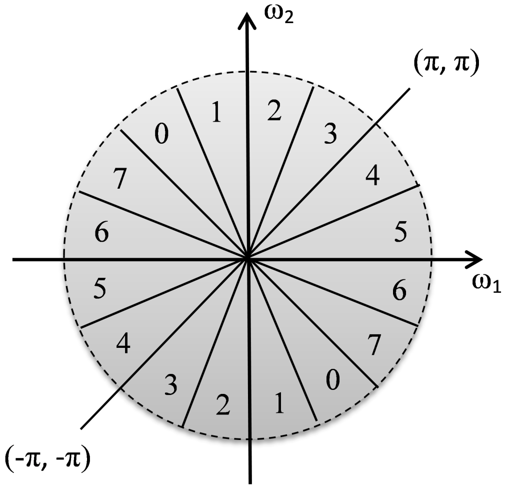


Figure 3.3: Directional Filter Bank. Frequency partitioning where $l = 3$ and there are 8 real wedge-shaped frequency bands. Subbands 0–3 correspond mostly to the horizontal directions, while subbands 4–7 correspond mostly to the vertical directions. The number of directions change with the value of l.

3.3.3 The Discrete Contourlet Transform

An important point to note is that the DFB is designed to capture only the high frequency components of the input image. This results in the leakage of the low frequency components into the directional sub-bands. This fact provides a reason to combine the DFB with a multi-scale decomposition where low frequencies of the input image are removed before applying the DFB. This is done using an iterative setup shown in figure 3.4 where the bandpass images from the LP are fed in to the DFB so that the directional information can be captured. The combined result is a double iterated filter bank structure named as a contourlet filter bank, which decomposes images into directional sub-bands at multiple scales.

The human beings as natural objects (i.e. not man-made) are made of a large collection of curved features e.g. edges. The CT that provides a near optimal representation of objects with curves is better suited to represent humans as compared to other transforms such as WT that has the limitation of using vertical and horizontal line segments to represent an edge structure.

3.4 Local Binary Patterns (LBP)

LBP belong to the class of non-parametric local image features that spatially exploit the geometric properties of a pattern and provides grey scale and rotation invariant texture features. The initial version of LBP introduced by Ojala et al. [106] labels the pixels of an image by thresholding, a 3×3 neighbourhood of each pixel with centre value, where 0 is assigned for a negative difference between the

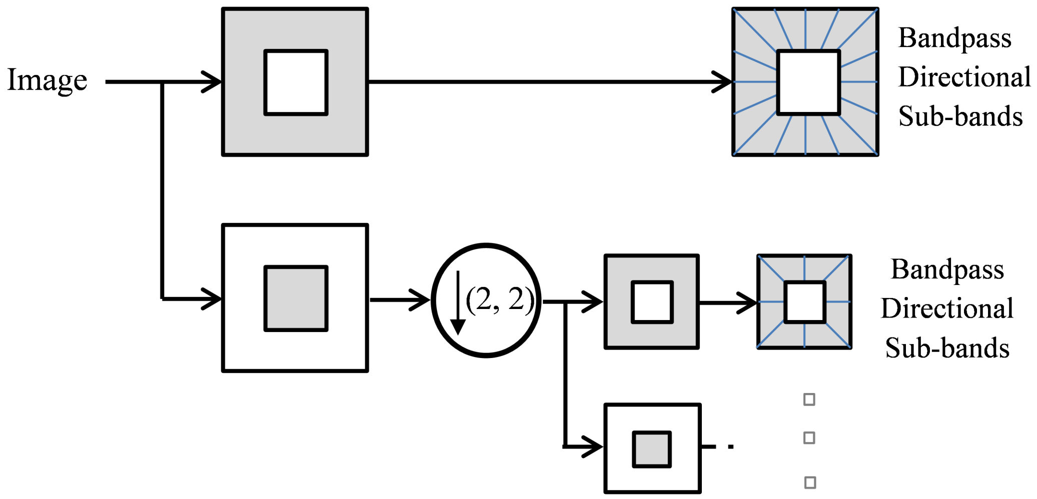


Figure 3.4: The countourlet filter bank: a multiscale decomposition is computed first using the LP and then a directional filter bank is applied to each bandpass channel.

centre pixel and the neighbouring pixel and 1 is assigned for positive difference. A histogram which gives the distribution of 256 binary patterns obtained from this procedure can be used as texture feature for subsequent analysis. The value of a $LBP_{P,R}$ code that takes P sample points with radius R around a pixel (x_c, y_c) is given by the following equation:

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(v_p - v_c)2^p, where, s(x) = \begin{cases} 1 & if \quad x \geq 0 \\ 0 & otherwise \end{cases} \tag{3.7}$$

where, v_c and v_p are the grayscale values of the centre pixel (x_c, y_c) and surrounding pixels (x_p, y_p). Figure 3.5 illustrates the process for the computation of LBP code around a pixel c.

3.4.1 Rotation Invariant LBPs

In a later work, Ojala et al. [105] extended the capability of the LBP operator for grey scale images and incorporated a rotation invariant property by considering a circular neighbourhood and bilinear interpolation. The rotation invariance of LBP ($LBP_{P,R}^{ri}$) is achieved by considering the minimum value of binary pattern. The minimum value is produced by shifting the binary structure to put maximum number of zeroes at the beginning of binary pattern. To remove the effect of rotation, i.e. to assign a unique identifier to each rotation invariant local binary

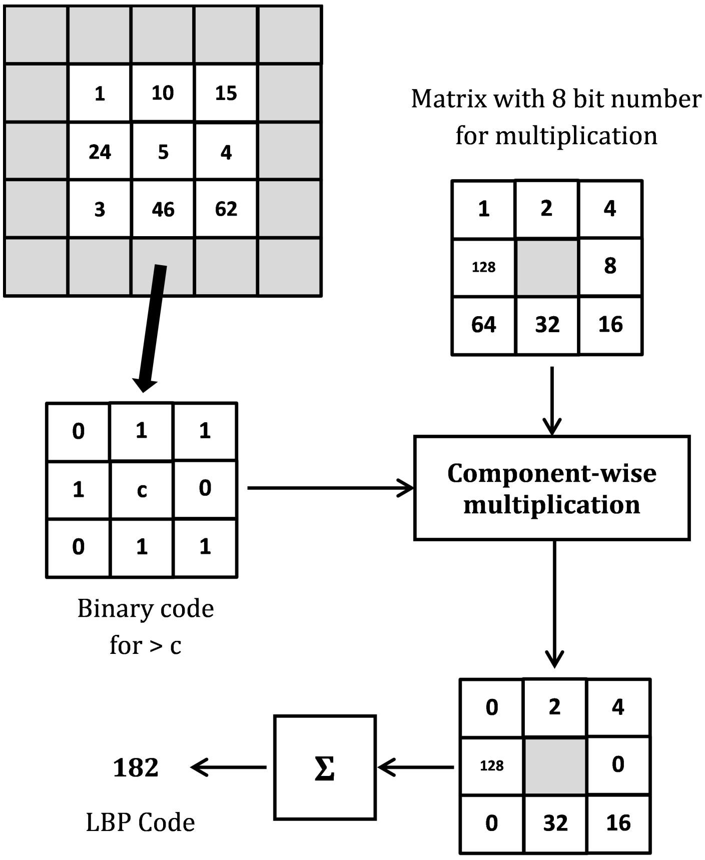


Figure 3.5: Computation of LBP code for pixel c with $P = 8$ and $R = 1$.

pattern, equation 3.8 is defined:

$$LBP_{P,R}^{ri} = \min\left\{\left(ROR(LBP_{P,R}, i) \,|i = 0, 1, ..., P - 1\right)\right\} \tag{3.8}$$

where ROR(x,i) performs a circular i times bit-wise right shift on the P-bit number x.

3.4.2 Uniform LBPs

Further in [105] it was observed that the binary patterns which have 0 to 1 or vice versa transitions are limited. For example, patterns 00111000, 11111111, 00000000, and 11011111 are uniform, and patterns 01010000, 01001110, or 10101100 are not. These patterns have been observed to capture most of the texture information. These patterns were named as uniform LBPs ($LBP_{P,R}^{riun}$) and mathematically represented in equation 3.9.

$$LBP_{P,R}^{riun} = \begin{cases} \sum_{p=0}^{p-1} s(v_p - v_c), & if \quad U(V_P) \leq n \\ P + 1 & if \quad otherwise \end{cases} \tag{3.9}$$

where,

$$U(V_P) = |s(v_{p-1} - v_c) - s(v_0 - v_c)| + \sum_{p=1}^{P-1} |s(v_p - v_c) - s(v_{p-1} - v_c)| \qquad (3.10)$$

Here, equation 3.10 measures the uniformity of the LBP code by summarizing the absolute value of the difference between the code and the code circularly shifted one bit.

The uniform LBP provide the reduced dimensionality of the feature vector. Hence, a small subset of the total of $2P$ patterns is sufficient to describe the texture of images. Here, n is the total number of binary patterns that exhibited the transitions. All codes that are classified as non-uniform are assigned with the value $P + 1$.

3.5 Colour Spaces

The choice of colour space is important while dealing with a particular application. Some colour spaces are similar to human perception of colours while others are devised for machines. A few well known colour spaces are discussed in the following subsections.

3.5.1 RGB

The RGB (Red-Green-Blue) representation is the most popular and is used in majority of imaging devices. It is an additive model where light is represented as the combination of three frequencies naming Red, Green and Blue. It is geometrically represented in the form of a cube (see figure 3.6) in which three primary axes are represented by R, G and B frequency components. The primary frequencies can be added to produce secondary colours which are:

- magenta = red + blue
- cyan = green +blue
- yellow = red + green

Examples of other possible combinations to produce white colour can be given as follows:

- white = blue (primary) + yellow (secondary)
- white = green (primary) + magenta (secondary)
- white = red (primary) + cyan (secondary)

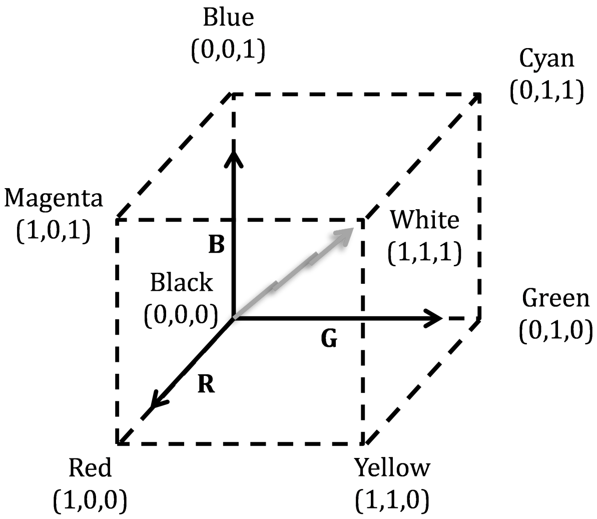


Figure 3.6: HSV and HSI colour spaces.

3.5.2 HSI and HSV

The HSI (Hue-Saturation-Intensity) space and HSV (Hue, Saturation and Value) space is obtained by non-linear transformation of the RGB space. The HSI representation uses the brightness (or intensity) value I = (R + G + B)/3 as the main axis orthogonal to the chrominance plane. The saturation S and the hue H are the radius and the angle, respectively, of the polar coordinates in the chrominance plane with the origin in the trace of the value axis (with R corresponding to 0^o) as shown in the figure 3.7. This representation is closely related to the way human vision perceives a colour because of its approximately perceptually uniform nature. Because of invariance to the object orientation with respect to illumination and camera viewing direction, the hue is more suitable for object retrieval. According to [151] the RGB to HSV(or I) colour space conversion is given as:

$$H = \cos^{-1}\left(\frac{0.5(R-G)+(R+B)}{\sqrt{(R-G)^2+(R-B)(G-B)}}\right) \tag{3.11}$$

$$S = 1 - \left(\frac{3}{R+G+B}\right)\min(R,G,B) \tag{3.12}$$

$$V(or I) = max(R,G,B) \tag{3.13}$$

Apart from RGB, HSV and HSI spaces, a number of other colour spaces have been presented in literature that have their significance in various applications. These colour spaces include YUV, YIQ, YDbDr, YCbCr, CMY(K), CIELab etc. The theoretical and mathematical background of these colour spaces is beyond the scope of this thesis therefore these colour spaces are not discussed here.

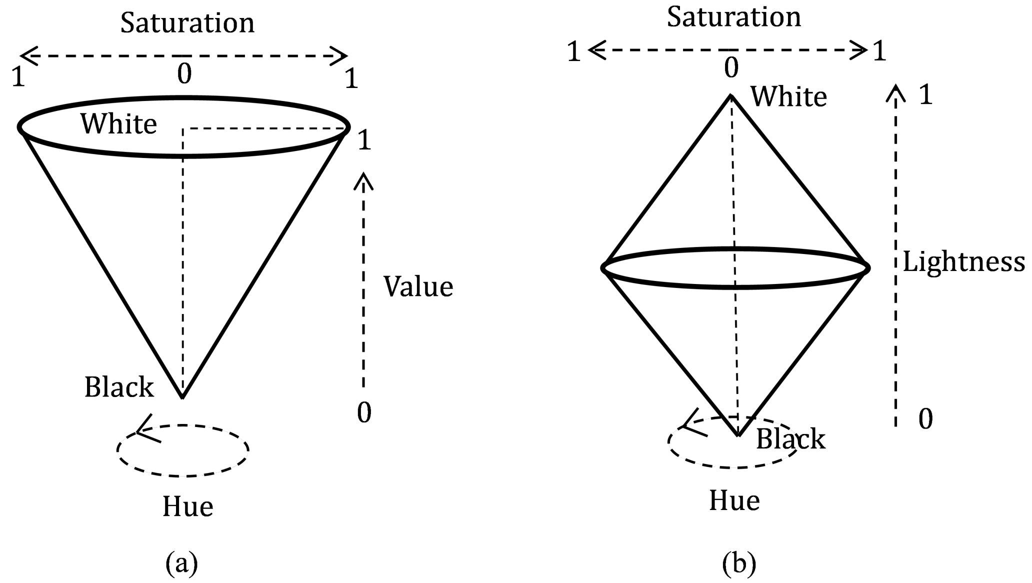


Figure 3.7: HSV and HSI colour spaces.

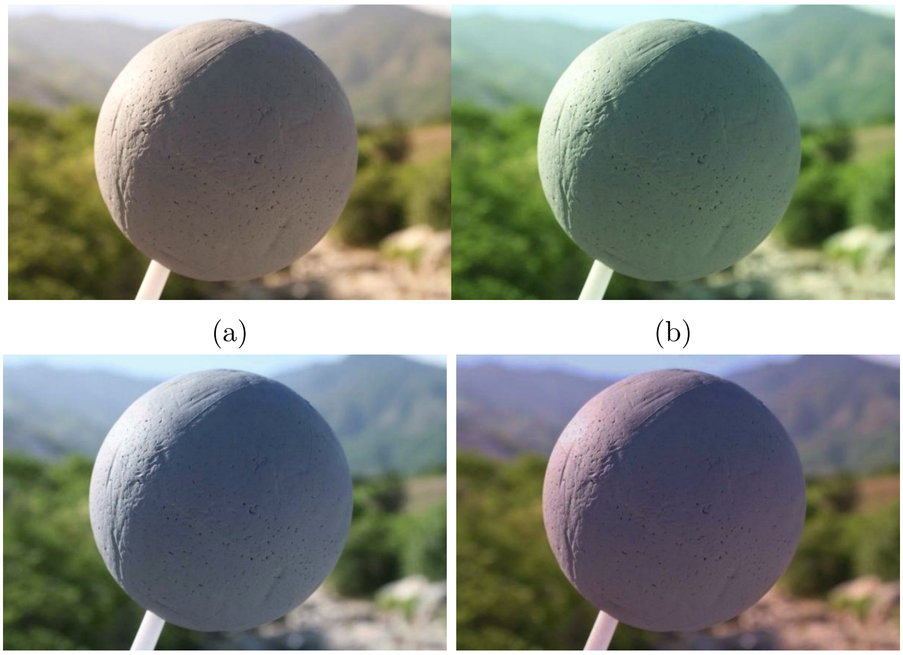
3.6 Colour Constancy

Colour Constancy is the ability to perceive colours of objects, invariant to the colour of the light source. This ability is generally accredited to the Human Visual System, although the exact details remain uncertain. An example of this human ability is shown in the figure 3.8. In this example, the same object is depicted four times, each rendered under a different light source. As can be seen, the colour of the object is strongly dependent on the colour of the light source.

Computational Colour Constancy can follow different paths to maintain a stable colour appearance across light sources. One common path, which does not mimic the human visual system, but is very common among computational models, approaches the problem using two phases. First, based on several assumptions, the colour of the light source (illuminant) is estimated from an input image. Then, using this estimated illuminant, the input image is corrected so that it appears to be taken under a canonical (e.g. white) light source.

3.6.1 Estimation of colour of the light source

One common assumption in literature on colour constancy is Lambertian shading, i.e. assuming that an image consists of only dull, matte surfaces. In this case, an image $f = (f_R, f_G, f_B)^T$ is composed of the multiplication of three terms, i.e. the camera sensitivity function $\rho(\lambda)$, the surface reflectance properties $S(x, \lambda)$ and the



(a) (b)

(c) (d)

Figure 3.8: Effect of different light sources on an image

colour of the light source $L(x, \lambda)$, then

$$f(x) = m(x) \int_{\omega} L(x, \lambda) S(x, \lambda) \rho(\lambda) d\lambda \tag{3.14}$$

where, ω is the visible spectrum, $m(x)$ is the Lambertian shading, λ is the wavelength of the light and x is the spatial coordinate in the image. Further assumptions include a spectrally uniform light source, i.e. $L(x, \lambda) = L(\lambda)$ for all locations x in the image. Then, the observed colour of the light source L depends on the spectrum of the light and the camera sensitivity function. The observed colour of the light source is expressed as:

$$L = \int_{\omega} L(\lambda) \rho(\lambda) d\lambda \tag{3.15}$$

The goal of colour constancy is to estimate L. Since, there are two unknown variables (the surface reflectance function $S(x, \lambda)$ and the colour of the light source L) and only one known variable (the image values f). Therefore, the estimation of L is an under-constrained problem. A common approach to solving this problem is to make further assumptions, for instance on the distribution of image colours or on the set of possible light sources.

3.6.2 Correction of Colours in the Images

Once the colour of the light source is known, the colour of the input image can be corrected. The transformation to convert an input image, recorded under an unknown light source, to an output image that appears to be recorded under a canonical light source, is called chromatic adaptation. Chromatic adaptation is often modelled using a linear transformation, which in turn can be simplified by a diagonal transformation when certain conditions are met. The diagonal model is explained here to correct the input images. The diagonal model is given by:

$$f_c = M_{u,c} f_u \tag{3.16}$$

where f_u is the image taken under an unknown light source, f_c is the same image transformed, so that it appears as if it was taken under a canonical light source, and $M_{u,c}$ is a 3×3 diagonal matrix that maps colours that are taken under an unknown light u to their corresponding colours under the canonical illuminant c. [11]

3.7 MPEG-7 Dominant Colour Descriptor

The Dominant Colour Descriptor (DCD) [26] is one of the seven colour descriptors defined by the MPEG-7 standard. It gives a description of the representative colours of an image. It is computed using the quantized colour histogram of an object. The computation of DCD not only eases the extraction of dominant colour but also allows the efficient indexing and retrieval of objects in large databases. The dominant colour descriptor is defined as:

$$D_{DC} = \{\{C_i, P_i, R_i\}, S\}, (i = 1, 2, ..., N) \tag{3.17}$$

where, C_i is a 3D dominant colour vector containing the component values of a colour space (H,S,V in this case), P_i is the percentage of each dominant colour and R_i is the colour variance that describes the variation of the colour values of the pixels in a cluster around the corresponding representative colour. The spatial coherence S is a single number that represents the overall spatial homogeneity of the dominant colours in the image and N is the number of dominant colours.

To measure the dissimilarity between to DCDs, following [135], the spatial coherence and the variance parameter can be neglected and the dissimilarity

$E(D_{DC1}, D_{DC2})$ can be expressed as:

$$E^2(D_{DC1}, D_{DC2}) = \sum_{i=1}^{N} P_{1i}^2 + \sum_{j=1}^{N} P_{2j}^2 - \sum_{i=1}^{N} \sum_{j=1}^{N} 2a_{1i,2j} P_{1i} P_{2j} \qquad (3.18)$$

where the subscripts 1 and 2 in all variables stand for descriptors D_{DC1} and D_{DC2} respectively, and $a_{k,l}$ is the similarity coefficient between two colours c_k and c_l and is computed using the following equation:

$$a_{k,l} = \begin{cases} 1 - \frac{d_{k,l}}{d_{\max}} & d_{k,l} \leq T_d \\ 0 & otherwise \end{cases} \qquad (3.19)$$

where $d_{k,l} = \|c_k - c_l\|$ is the Euclidean distance between two colour clusters c_k and c_l. The threshold T_d is the maximum distance used to judge whether two colour clusters are similar and $d_{max} = \alpha T_d$, where α is used to parametrize the maximum distance T_d between two colour clusters.

3.8 Bilateral Filtering

One of the most fundamental operations of signal processing is the Filtering. In terms of image processing, the filtering process constitutes a procedure that replaces the values at any given location in an image with the values or function of those values of a small neighbourhood of that location. The noise values that corrupt these nearby pixels are mutually less correlated than the signal values, so noise is averaged away while the signal is preserved.

The assumption of slow spatial variations does not hold at edges, which are consequently blurred by linear low-pass filtering. Many efforts have been devoted towards reducing this undesired effect of averaging across edges. One of them is Bilateral Filtering (BF) [152], it is a simple, non-iterative scheme for edge-preserving smoothing.

According to [152], the basic idea underlying BF is to apply those operation to the range of an image what traditional filters apply to its domain. Two pixels can be close to one another, that is, occupy nearby spatial location, or they can be similar to one another, that is, have nearby values, possibly in a perceptually meaningful fashion.

Consider a translation-invariant low-pass domain filter $g(\vartheta, x)$ applied to an image $I(x)$:

$$h(x) = k_d^{-1}(x) \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} I(\vartheta) g(\vartheta, x) d\vartheta \qquad (3.20)$$

where, $g(\vartheta, x)$ measure the geometric closeness between the neighbourhood centre x and a nearby point ϑ. If the low-pass filtering is to preserve the DC component of low-pass signals, then,

$$k_d(x) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(\vartheta, x)d\vartheta \tag{3.21}$$

If the filter is translation-invariant, $g(\vartheta, x)$ is only a function of vector difference $\vartheta - x$, and k_d is constant. Range filtering is similarly defined as:

$$h(x) = k_r^{-1}(x)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(\vartheta)p(I(\vartheta) - I(x))d\vartheta \tag{3.22}$$

except that now $p(I(\vartheta)-I(x))$ measures the photometric similarity between the pixel at the neighbourhood centre x and that of a nearby point ϑ. The closeness function g operates in the domain of the image function I while the similarity function p operates in the range of I. The normalization constant in this case is:

$$k_r(x) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(I(\vartheta) - I(x))d\vartheta \tag{3.23}$$

The spatial distribution of image intensities plays no role in range filtering taken by it. Combining intensities from the entire image, however, makes little sense, since the distribution of image values far away from x ought not to affect the final value at x. In addition, one can show that range filtering without domain filtering merely changes the colour map of an image, and is therefore of little use. It is important to enforce both geometric and photometric locality by combining domain filtering and range filtering. Combined filtering can be described as follows:

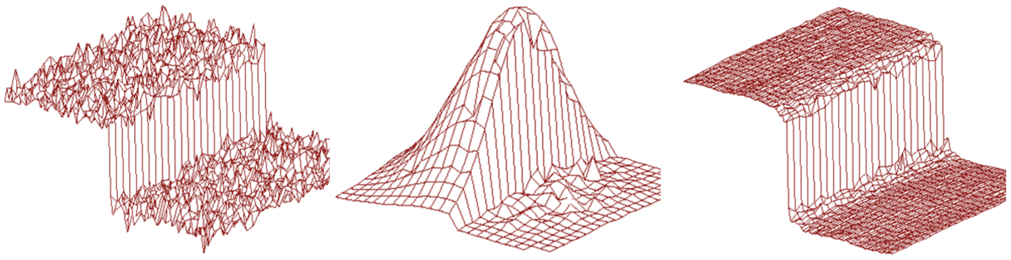
$$I_{sm}(x) = k^{-1}(x)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(\vartheta)g(\vartheta, x)p(I(\vartheta), I(x))d\vartheta \tag{3.24}$$

with the normalization:

$$k(x) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(\vartheta, x)p(I(\vartheta), I(x))d\vartheta \tag{3.25}$$

Combined domain and range filtering is referred as BF. It replaces the pixel value at x with an average of similar and nearby pixel values. In smooth regions, pixel values in a small neighbourhood are similar to each other, and the BF acts essentially as a standard domain filter, averaging away the small, weakly correlated differences between pixel values caused by noise.

Consider now a sharp boundary between a dark and a bright region, as in figure 3.9. When the bilateral filter is centred, say, on a pixel on the bright side of the boundary, the similarity function assumes values close to one for pixels on the same side, and values close to zero for pixels on the dark side. The similarity



(a) (b) (c)

Figure 3.9: (a) A 100-gray level step perturbed by Gaussian noise with $\sigma = 100$ gray levels. (b) Combined similarity weights for a $23x23$ neighbourhood centred two pixels to the right of the step in (a). The range component effectively suppresses the pixels on the dark side. (c) The step in (a) after bilateral filtering with $\sigma_r = 50$ gray levels and $\sigma_d = 5$ pixels. [152]

function is shown in figure 3.9b for a 20×20 filter support centred two pixels to the left of the step in figure 3.9a. The normalization term $k(x)$ ensures that the weights for all the pixels add up to one. As a result, the filter replaces the bright pixel at the centre by an average of the bright pixels in its vicinity, and essentially ignores the dark pixels. Conversely, when the filter is centred on a dark pixel, the bright pixels are ignored instead. Thus, as shown in figure 3.9c, good filtering behaviour is achieved at the boundaries, due to the domain component of the filter, and crisp edges are preserved at the same time, due to the range component.

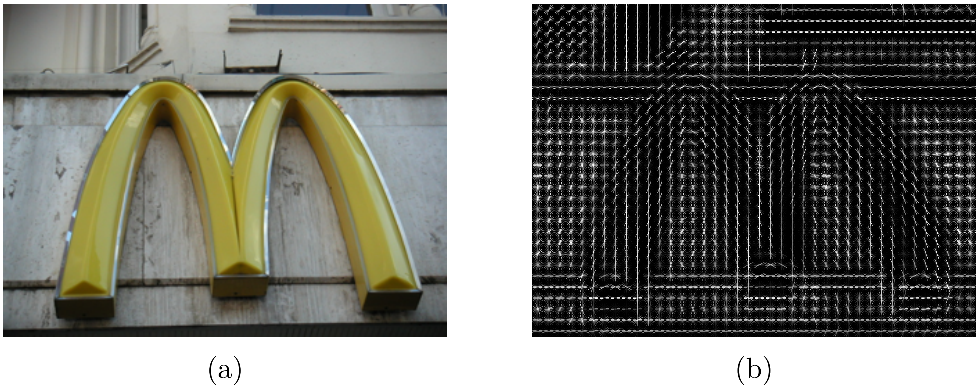
3.9 Histogram of Oriented Gradients

The Histogram of Oriented Gradients features were presented by Dalal and Triggs [32]. The method exploits the gradient information present in the image and finds the local histograms of gradient orientations in small sub-windows of the image.

The process of HOG computation starts by computing the local gradient information in the intensity image. The image is then divided into small spatial regions (named as cells) of equal size. For each cell, a 1-D local histogram is computed from the gradient direction of every pixel. The final HOG representation is computed by concatenating the histogram from all the cells of the image. Figure 3.10 shows an input image and the corresponding HOG features with a cell size of 8 pixels.

3.10 Scale Invariant Feature Transform (SIFT)

Scale Invariant Feature Transform (SIFT) is a well known approach for detecting the key-points and extracting feature descriptors around those key-points within



(a) (b)

Figure 3.10: HOG Features. (a) Original Image. (b) Standard HOG features with a cell size of eight pixels.

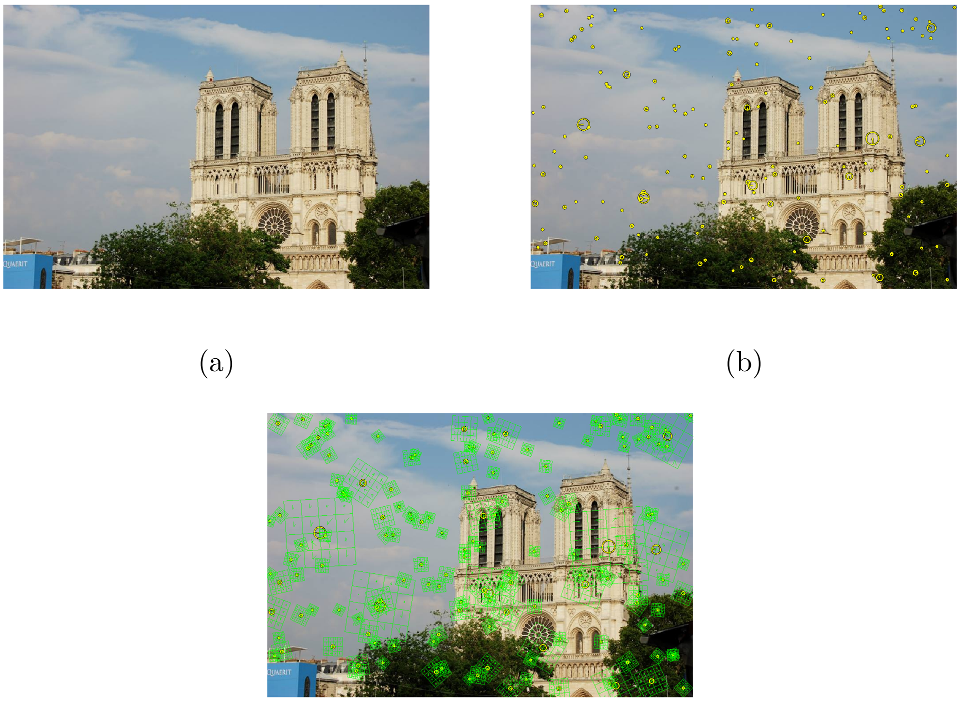
the image. The SIFT description is highly discriminative and offers the advantage of being fully invariant to scale and rotation and partially invariant to affine distortion, noise, viewing angle and illumination. There are four key stages in the computation of SIFT given as:

Detection of Scale-space Extrema: The first step is the detection of key-points within the image. The key-points are basically high frequency points in the image such as corners. In order to do that, scale-space filtering such as Laplacian of Gaussian is used with various values of σ, where σ is the standard deviation of Gaussian distribution. The reason being the small values of σ fits well for small corners while large values of σ fits for larger corners. Eventually, the local maxima across the scale and space is computed giving a list of (x, y, σ) with (x, y) being the coordinates of the potential key-points with scale σ.

Keypoint Localization: The potential key-points detected in the previous stage are further analysed using Taylor series expansion of scale space to get accurate location of extrema. If the intensity at a particular extrema is less than a certain threshold then it is rejected. The process eliminates any low-contrast key-points that were detected in the previous step. Figure 3.11b shows a few localized key-points in the image.

Orientation Assignment: In order to achieve invariance to image rotation, an orientation is assigned to each key-point. This is done by selecting a neighbourhood around the key-point location based on the scale and the gradient magnitude. The direction in that region is computed using an orientation histogram with 36 bins covering 360 degrees. The gradient magnitude and gaussian weighted circular window with σ equal to 1.5 times the scale of the key-point is used as a weighting factor for orientation histogram. The highest peak in the histogram is along with any peaks above 80 % of it is considered to calculate the orientation. This results in key-points with same location and scale but different directions.

Keypoint Descriptor: The descriptor around each key-point is computed by selecting a 16 × 16 neighbourhood around it (see figure 3.11c). It is divided in to 16 sub-blocks of 4 × 4 size. An 8 bin orientation histogram is created by each sub-block. The histograms from all sub-blocks are concatenated to form a 128 dimensional feature descriptor.



(a)

(b)

(c)

Figure 3.11: Computation of SIFT. (a) Original Image. (b) Some of the detected key-points. (c) Descriptor window around each key-point.

Dense SIFT: The dense SIFT features on an image is computed using a dense gird of locations at a fixed scale and orientation. This type of feature descriptors is often uses for object categorization. This thesis exploits the dense SIFT features to capture the pattern within the frontal vehicle images.

3.11 Local Energy based Shape Histogram

The local energy signifies the underlying corners, edges or contours in an image. In [125] the authors made use of local energy information to encode shapes and proposed a histogram based representation of the local energy along various orientations. The theoretical aspect of their technique is explained in this section.

The local energy model was first proposed by Morrone and Owen in [95]. It postulated that features are perceived in an image at points where the local frequency components have maximum phase congruency.

$$E(x) = max_{\bar{\varphi}(x)\in[0,2\pi]} \frac{\sum_n A_n \cos(\varphi_n - \bar{\varphi}(x))}{\sum_n A_n} \tag{3.26}$$

Where A_n and φ_n are the magnitude and phase component of the nth Fourier component. This frequency information is obtained using pairs of symmetric and anti-symmetric linear filters to preserve the underlying phase information. This is done by convolving the image with a filter bank containing Gabor wavelet kernels tuned to 5 spatial frequencies and 8 orientations. The operation is mathematically expressed as:

$$G(e_{n,v}, o_{n,v}) = I(x,y) * \psi_{n,v}(z) \tag{3.27}$$

where $G(.)$ is the response of Gabor filter bank ψ at scale n and orientation v at location (x, y) of image I. The amplitude A_n and phase ϕ_n of the response is computed using the following equation:

$$\begin{cases} A_n = \sqrt{e_n^2 + o_n^2} \\ \phi_n = \tan^{-1}\left(\frac{e_n}{o_n}\right) \end{cases} \tag{3.28}$$

The modified energy model presented in [73] consists of sine of the phase deviation, including a proper weighing of the frequency spread W and also a noise cancellation factor T. The normalization by summation of all component amplitudes makes it independent of the overall magnitude of the signal, making it invariant to illumination variations in images. The local energy analysis is intended to detect interest points in images with a high reliability in presence of noise and illumination. The local energy, as defined by Kovesy in [73] is given as:

$$E = \frac{\sum_n W(x) \left\lfloor A_n(x)(cos(\phi_n(x) - \bar{\phi}(x)) - \left|\sin(\phi_n(x) - \bar{\phi}(x))\right| - T \right\rfloor}{\sum_n A_n(x) + \varepsilon} \tag{3.29}$$

The energy model given in equation 3.26 has been used to generate a local histogram accumulating the local energy along each filter orientation on n subregions of the image [125]. An orientation label map L is computed by assigning each pixel the label of the orientation at which the local energy is maximal across all scales. A local histogram of the energy is accumulated along each filter orientation for each sub region. The local histograms h are computed according to the following equation:

$$h_{r,b} = \sum w_r \times E \times \delta_{Lb} \tag{3.30}$$

where δ_{Lb} represents the Kronecker delta, E is the local energy, b is the current bin and w is Gaussian weighting function centred at region r. The weight function is computed as:

$$w_r = \frac{1}{2\pi\sigma} e^{\frac{\left[(x-r_{xo})^2+(y-r_{yo})^2\right]}{\sigma^2}} \tag{3.31}$$

The image is partitioned into 16 blocks and 8-bin local histogram corresponding to 8 filter orientations is computed on each block. The histograms on all image partitions are concatenated to maintain spatial relationship between the content of the image, making it a 128 dimensional feature vector.

3.12 Feature Encoding Techniques and the Fisher Vector (FV)

The purpose of using an encoding scheme is to achieve discriminative dimensionality reduction in the case of dense feature extraction. Various encoding schemes i.e. fisher, histogram, kernel codebook, super vector etc. is in use at present and research is under way to employ and improve these strategies to work efficiently for discriminative dimensionality reduction. Fisher encoding [64] captures and averages the first and second order difference between the image descriptor and the centres of a Gaussian Mixture Model (GMM), which may be thought of as a soft visual vocabulary.

As mentioned in the literature review chapter, a number of feature encoding schemes have recently emerged. In [20], the authors have elaborated the superiority of FV among modern schemes.

Perronnin et al. [114] devised a modified FV representation keeping image classification domain under consideration. The set of descriptors is modelled with a GMM using pre-specified number of Gaussian distributions. The FV representation is achieved by computing the average or first and second order differences between image descriptors and GMM centres.

Let $I = (x_1, ..., x_N)$ be a set of D dimensional feature vectors (SIFT descriptors in this case) extracted from a key-patch. Let $\Theta = (\mu_k, \Sigma_k, \pi_k : k = 1, ..., K)$ be the parameters of a Gaussian Mixture Model fitting the distribution of descriptors, where μ_k, Σ_k, π_k is the mean, variance and weight respectively of kth gaussian distribution in the GMM. The GMM associates each vector x_i to a mode k in the mixture with a strength given by the posterior probability:

$$s_{ik} = \frac{e^{\left(-\frac{1}{2}(x_i - \mu_k)^T \sum_k^1 (x_i - \mu_k)\right)}}{\sum_{t=1}^{K} e^{\left(-\frac{1}{2}(x_i - \mu_t)^T \sum_k^{-1} (x_i - \mu_t)\right)}} \tag{3.32}$$

For each mode k, the mean and covariance deviation vectors α_k and β_k respectively are given as:

$$\alpha_{jk} = \frac{1}{N\sqrt{\pi_k}} \sum_{i=1}^{N} s_{ik} \left(\frac{x_{ji} - \mu_{jk}}{\sigma_{jk}} \right) \tag{3.33}$$

$$\beta_{jk} = \frac{1}{N\sqrt{2\pi_k}} \sum_{i=1}^{N} s_{ik} \left[\left(\frac{x_{ji} - \mu_{jk}}{\sigma_{jk}} \right)^2 - 1 \right] \tag{3.34}$$

where, $j = 1, 2, ..., D$ spans the vector dimensions. The FV of image I is achieved by concatenating all the computed difference vectors α_k and β_k for each of the K modes in the Gaussian mixtures:

$$FV = [\alpha_1, \beta_1, ..., \alpha_K, \beta_K] \tag{3.35}$$

The FV representation captures the discriminative information present in the densely computed feature descriptors of an images.

3.13 Support Vector Machine (SVM)

A Support Vector Machine (SVM) [157] is a technique to train a discriminative classifiers. The classifier is formally defined by a separating hyperplane. In simple words, given a set of labelled training data, the SVM outputs an optimal hyperplane which categorizes new examples. One of the main advantages of SVM is its capability of learning the data pattern in a high-dimensional space even with a very few number of training examples. At the same time, it minimizes a bound on the empirical error and the complexity of the classifier. This concept is formalized in the theory of uniform convergence in probability using the equations 3.36. For a probability η such that $0 \leq \eta \leq 1$, the losses as explained in equation 3.37 taking these values, with probability $1 - \eta$, the following bound holds:

$$R(\alpha) \leq R_{emp}(\alpha) + \Phi\left(\frac{h}{l}, \frac{-\log(\eta)}{l}\right) \tag{3.36}$$

where $R(\alpha)$ is the expected risk, $R_{emp}(\alpha)$ is the empirical risk, l is the number of training examples, h is the dimension of the classifier that is being used and is a non-negative integer named as Vapnik Chervonenkis (VC) dimensions, and $\Phi(.)$

is the confidence of the classifier.

The empirical risk $R_{emp}(\alpha)$ is defined to be the measured mean error rate on the training set (for a fixed, finite number of observations):

$$R_{emp}(\alpha) = \frac{1}{2l} \textstyle\sum_{i=1}^{l} |y_i - f(x_i, \alpha)| \tag{3.37}$$

Here, the quantity $1/2l|y_i - f(x_i, \alpha)|$ is called the loss. Note that no probability distribution appears here. $R_{emp}(\alpha)$ is a fixed number for a particular choice of α and for a particular training set x_i, y_i.

Intuitively, what this means is that the uniform deviation between the expected risk and empirical risk decreases with larger amounts of training data and increases with the dimension, h. This leads directly to the principle of structural risk minimization, whereby an attempt can be made to minimize at the same time both the actual error over the training set and the complexity of the classifier; this will bound the generalization error as in equation 3.36. It is exactly this technique that SVMs approximate. This controlling of both the training set error and the classifier's complexity has allowed SVMs to be successfully applied to very high dimensional learning tasks; [68] presents results on SVMs applied to a 10,000 dimensional text categorization problem and [107] show a 283 dimensional face detection system. The two major types of SVMs are: Linear SVM and Non-Linear SVM. For in-depth details and mathematical formulation of both type of SVM, the reader is referred to [158] and [157].

3.14 Cosine Similarity (CS)

The Cosine Similarity (or Distance) is a measure that calculates the cosine of the angle between two vectors. The metric is the measurement of the orientation irrespective of any difference between the magnitudes of vectors. CS measure has proved to be very effective for computing the similarities and the differences between patterns represented by feature descriptors. The CS $C(a, b)$ between two vectors a and b can be computed using the following equation:

$$C(a, b) = \cos(\theta) = \frac{a \bullet b}{||a|| \; ||b||} \tag{3.38}$$

Using the above equation, the Cosine Distance $C^{'}(a, b)$ between two vectors a and b can be computed as:

$$C^{'}(a, b) = 1 - C(a, b) \tag{3.39}$$

3.15 Summary and Conclusion

This chapter discussed the major concepts and methodologies that have been used to devise the frameworks proposed in contributory chapters (4-7). The background modelling algorithm described in section 3.2 has been used in chapter 4 to develop a more robust version of the method particularly for surveillance video applications. The combination of HOG features (section 3.9) and SVM (section 3.13) based classification technique has been used as a baseline to compare the moving object recognition framework in chapter 4. The HOG features have been significantly used in the text detection and recognition work in chapter 7 to represent candidate regions for text detection and to represent character images in recognition task. The colour constancy techniques explained in section (3.6) and the concepts regarding the conversion of colour spaces explained in section (3.5) are directly used in the proposed frameworks for the correction and recognition of colours in chapter 5 and for the detection and recognition of text information in chapter 7. Further, the dense SIFT features (section 3.10) and FV encoding algorithm (3.12) is the backbone of the VMMR technique presented in chapter 6. It was therefore important for the better understanding of the reader to briefly review the feature encoding mechanism. In addition to that, LESH feature extraction mechanism is also discussed is section 3.11 as its efficiency has been analysed in the proposed vehicle make and model recognition framework in chapter 6.

Chapter 4

Moving Object Detection and Recognition

4.1 Introduction

Moving object detection and recognition is the first and most important step in video content analysis. In terms of moving targets, humans and vehicles are the two salient objects that appear in surveillance videos. Foreground segmentation and object recognition are two different problems. However, if the candidate regions for object recognition are extracted through foreground segmentation technique then the performance of object recognition directly depends on the accuracy of foreground segmentation technique. In such scenarios, an accurate foreground extraction ensures better object recognition performance. This chapter addresses both problems together to address the challenges involved in the annotation of surveillance videos.

In terms of moving object detection and foreground extraction, a novel pipeline has been devised to improve the performance of existing foreground segmentation algorithm of [148] which is sensitive to background clutter and often the output is significantly disturbed in the presence of shadow. These factors degrade the performance of foreground extraction algorithm and pose a hurdle in the accurate extraction of contours of moving targets especially under varying illumination conditions and in the presence complex background. The proposed modification enhances the robustness and accuracy of the algorithm [148] in the presence of various factors such as noise that appears in the form of background clutter and shadows etc. The experiments conducted on benchmark datasets show that the proposed technique achieve better performance as compared to baseline technique (see section 4.5).

In object recognition, the authors in [72] use a combination of Wavelet Transform (WT) and Local Binary Patterns (LBP) features for pedestrian detection. However, a potential problem in the WT is its inability to highlight the contour of the objects that possess curves in their silhouette (e.g. humans). The filters in the WT filter bank capture only those pixels that strictly possess the orientation that the filter is trained to capture. This results in the separation of adjacent pixels into various WT coefficients, consequently the texture in the silhouette is not captured in any of the WT coefficients. The Contourlet Transform (CT) on the other hand provides the advantage that it captures the contours instead of pixels or straight lines. Therefore, it can capture the silhouette of objects such as humans that possess curves in their contours. This provide the basis for the development of a framework to exploit the advantage of CT.

A novel feature descriptor using CT and Centre Symmetric Local Binary Patterns (CS-LBP) features is proposed and exhaustively tested for its performance in classification of two main subjects: humans and vehicles. The proposed feature descriptor exploits the information present in the contours of the foreground objects. A model of each object class is implicitly derived from the feature representation of the training set in an example-based learning approach. A Support Vector Machine (SVM) based learning approach has been followed here to derive the discriminative model of each object class. The proposed feature descriptor is closely related to the one presented by Pan et al. [108], however, unlike [108] the proposed framework analyse the CT coefficients to deeper levels of decomposition to capture the most relevant texture while suppressing the noise. The work presented in this chapter is based on the references [187] and [49].

4.2 Foreground Extraction

The first step in the analysis of moving objects in videos is the segmentation of these objects from background. A modified version of the foreground segmentation method of [148] is devised in this work to reduce the shortcomings present in the original version. The weaknesses in Stauffer and Grimson's [148] method includes the inability to deal with shadows, spurious objects, fast background initialization, illumination changes and similarities in foreground-background regions. This led to the proposal of a number of new approaches ([166, 131, 39, 165, 66]) aiming at improving the original concept.

The proposed foreground segmentation process starts by modelling the background using a Mixture of Gaussian (MoG) distributions. Each pixel in the video frames is modelled by a weighted MoG the parameters and weights of which are continuously updated with every new frame. This allows the prediction of the pixel values that belong to the background in consecutive frames based on prior prob-

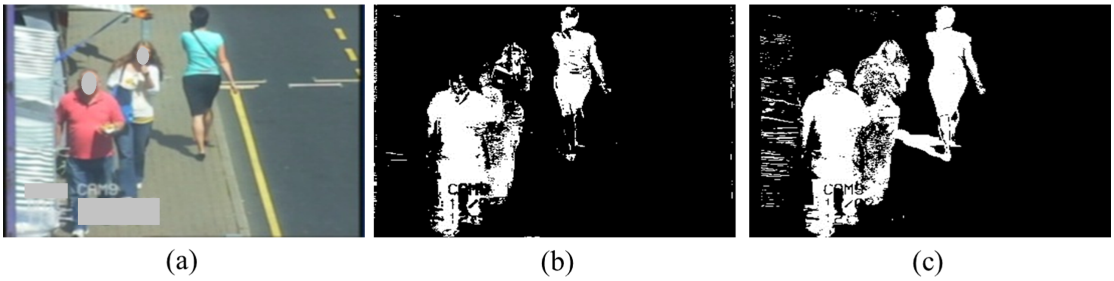


Figure 4.1: The effect of shadow removal on foreground segmentation. (a) Original video frame (b) Extracted foreground regions along with shadow removal based on intensity information of normalized RGB image. (c) The result of MoG based foreground extraction on RGB image.

ability. This work combines a few of the successful improvements of the method in [148] to address the aforementioned issues in order to utilize the Gaussian Mixture Model (GMM) while overcoming its weaknesses.

The most common approaches adopted by researchers [166, 131, 39, 165] to eliminate shadows are based on intensity analysis. All the techniques suggest applying the Stauffer-Grimson's algorithm in the normalized R_nG_nI colour space, where I is the intensity component calculated as the average of the Red (R), Green (G) and Blue (B) components. The pixels belonging to shadow regions and highlighted regions are then classified by setting up a threshold on the ratio $r = I_c/I_m$, where I_c is the intensity of the current frame and I_m is the mean intensity over the sequence of all previous frames. The use of normalized RBG space generates good performance in various scenarios especially where the background does not contain similar colour as shadow regions but the algorithms badly fails in scenarios where foreground contains coloured pixels similar to shadow regions. It has been found out with extensive experimentation that the use of normalized colour space (R_nG_nI) degrades the performance of foreground extraction by the removal of such foreground pixels that have low intensity values (i.e. pixels in grey or dull white colour regions). Figure 4.1b illustrates this problem where some parts of the foreground regions have been eliminated along with shadow regions. In order to avoid the rejection of pixels in target Regions of Interest (RoI), it has been empirically observed that the use of RGB instead of normalized RGB reduces the rejection of RoI pixels. Figure 4.1c illustrates this mechanism where all foreground pixels have been successfully separated from background along with a few noise pixels.

Inspired by the work of Javed et al. [66], this work exploits the gradient information in the image to recover missing foreground region while eliminating the pixels belonging to shadow regions and other spurious objects of the background. Similar to [66], first the Stauffer-Grimson's algorithm is applied on the gradient

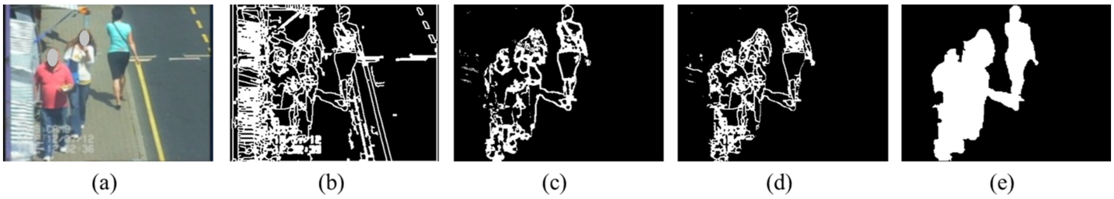


Figure 4.2: The illustration of accurate contour definition of the foreground regions. (a) Original video frame. (b) Dilated edge map of the video frame. (c) Foreground regions extraction by background subtraction on the gradient images. (d) Edges corresponding to foreground regions. (e) The Output, showing foreground binary map with well-defined contours.

magnitude image I_g to acquire the foreground edge map. The foreground edge map is denoted by $E_{forg}(I_g)$. Figure 4.2c shows the edge map $(E_{forg}(I_g))$ of the image in 4.2a. Next, a similar region level processing to [66] is applied to address the problem of varying illumination and elimination of spurious objects. The work of 4.2 addresses foreground extraction under varying illumination, however, it does not have the capability to deal such scenarios where moving targets have shadows along with them. The output deteriorates when moving subjects are not well separated and possess shadows. The objects connected because of shadows are considered as single object by the algorithm.

The advantage of proposed method over [66] is that it exploits the edge map of extracted foreground objects $(E_{forg}(I_g))$ to remove regions belonging to shadows of those objects while robustly handling varying illumination in the scene.

The primary goal of the proposed approach is the accurate mapping of the contours of foreground regions. Let, the original coloured frame is I and the foreground region acquired using Stauffer-Grimson's method [148] be $F(I)$ and the dilated binary edge map of the image I be $E(I)$. A post processing stage comprises of morphological dilation and closing operations is applied to achieve the continuous edges in the edge maps $E_{forg}(I_g)$ and $E(I)$.

At the next stage, a convex hull is computed to enclose $E_{forg}(I_g)$. The convex hull is used as a reference to discard all such pixels of $F(I)$ that fall out of the computed convex hull; thus achieving the elimination of such regions that do not possess strong gradient values (e.g. shadows). The target contour region C comprises of all those pixels of $E(I)$ that are resent in the sum of $F(I)$ and $E_{forg}(I_g)$. Mathematically, it can be expressed using the following equation:

$$C = E(I)\left[F(I) + E_{forg}(I_g)\right] \tag{4.1}$$

The pixels enclosed in the contour C are considered as the foreground pixels as illustrated in figure 4.1e. Since the extracted contours have minor irregularities and breakages, a post processing stage involving Median filtering is applied to the detected foreground binary map in order to acquire smooth edges in the foreground binary map.

The main advantage of the proposed pipeline is its ability to achieve well-defined contours of the foreground regions. The conventional MoG based algorithms fail to extract an accurate and well defined contours of the foreground region. On the other hand, the convex hull based approach alone does not always enclose the whole object that results in broken and irregular foreground patterns and also results in gaps between human lower limbs and the torso to be detected as foreground.

The extracted foreground binary map using proposed pipeline can be directly used to identify and annotate the moving objects on the basis of appearance. However, in cases where more than one object are moving together as shown in the figure 4.2a, the extracted foreground binary map of both objects accumulates into single binary map that cannot be used to identify that object. The coordinates of the bounding box enclosing the foreground binary map are used to extract the window from the original video frame to perform object identification. The main reason to include the foreground extraction stage is to exploit temporal information to find the location of the potential objects of interest i.e. humans and vehicles. This condenses the overall search area only to those parts of the frame where an object is detected unlike the conventional object detectors where sliding window based search is performed in the entire video frame.

4.3 Object Recognition

The bounding boxes of the moving objects that are detected in the previous step are subsequently passed to the object recognition pipeline. The recognition framework classifies the object in the bounding box window. A novel feature descriptor is devised to fulfil the task of object recognition. The feature descriptor of the foreground window is used to classify it to one of the objects of interest (human or vehicle) or the region belonging to neither of them (i.e. background). Note that it is not within the scope of this work to perform recognition of those objects that are partially occluded.

4.3.1 Countourlet Transform based Centre Symmetric Local Binary Patterns (CCS-LBP)

As mentioned in the literature review chapter, the authors in [190] and in [108] presented texture analysis of frequency coefficients based feature descriptor for human detection and for object classification respectively; this work is inspired by their findings. The proposed work focuses mainly on the classification of moving objects i.e. humans and vehicles. The human body in standing position has strong vertically oriented contours and edges along its boundaries. In the case of vehicles, it has been observed that the orientation of contours and edges along both horizontal and vertical axis is approximately the same. The edge map of the boundary of the vehicles contains long horizontal and vertical edges due to the geometric shape of the vehicles. It is also observed that the spatial structure of the human body and a vehicle possess a bilateral symmetry. Therefore, the detail in the contour is the most relevant feature for identifying human and vehicle objects.

The LBPs are suitable for modelling repetitive textures which means that these features are sensitive to random noise in uniform image areas. However, the extraction of LBP in CT domain helps to analyse the contours of the objects while reducing the effect of noisy patterns that are present within in the body of these structures and in the surrounding background. These patterns influence the pattern of interest and acts as noise in the computed feature descriptor.

In this study, a variation of conventional LBP technique called the Center Symmetric Local Binary Patterns [190] is exploited to represent the pattern in the CT sub-bands. The CS-LBP technique compares only center symmetric pairs of pixels to produce more compact binary patterns. The LBP produces 256 different binary patterns for 8 neighbours whereas for CS-LBP this number is only 16. This offers a significant advantage in terms of the reduced dimensionality of the feature descriptor. The CS-LBP is given as:

$$CS{-}LBP_{R,N,T}(x,y) = \sum_{i=0}^{({}^{N}\!/_{2})-1} s\left(n_i - n_{i+({}^{N}\!/_{2})}\right)2^i, s(x) = \begin{cases} 1 & x > T \\ 0 & otherwise \end{cases} \tag{4.2}$$

where n_i and $n_{(i+(N/2))}$ corresponds to the grayscale values of centre-symmetric pairs of N equally spaced pixels on a circle of radius R.

The CT is used here as a multi-scale and multi directional global image contour representation scheme. The extraction of LBP on a set of selective coefficients of CT enables to capture the contour pattern, which can be effectively used to identify the object. The proposed feature representation scheme is named as Contourlet based Center Symmetric Local Binary Patterns (CCS-LBP). The proposed feature descriptor captures the pattern in the silhouette of the objects for discriminative representation of their shape.

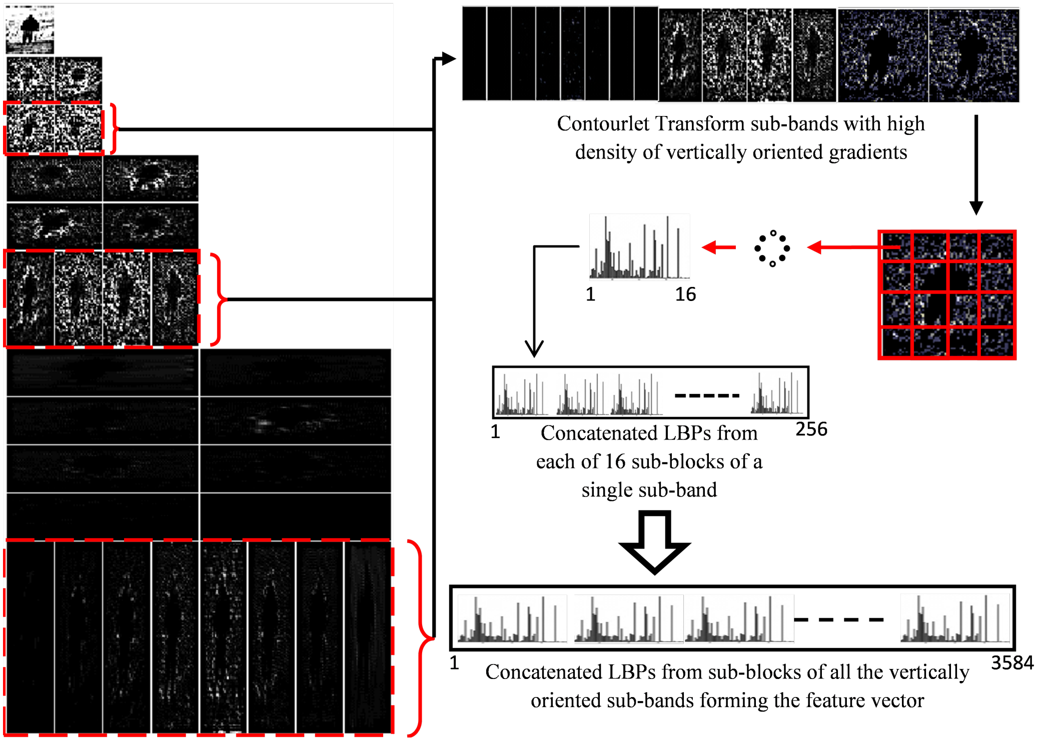

Contourlet Transform sub-bands with high density of vertically oriented gradients
1
16
1
Concatenated LBPs from each of 16 sub-blocks of a single sub-band
256
1
Concatenated LBPs from sub-blocks of all the vertically oriented sub-bands forming the feature vector
3584

Figure 4.3: Computation of CCS-LBP feature descriptor.

4.3.2 Extraction of Feature Vectors

The CT coefficients are computed for every foreground window $W(x, y)$ as a first step in the classification stage. The window contains foreground object along with some part of the background. The number of decomposition levels l and the number of directional decompositions d at each scale are pre-specified. The empirically selected values in the experiments are $l = 3$ and $d = [4, 8, 16]$ that results in 28 sub-band coefficients (see figure 4.3). The magnitude and sign of each CT sub-band coefficient $C_{l\_d}(x, y)$ according to [108] is given as:

$$mag[C_{l\_d}(x, y)] = |C_{l\_d}(x, y)| \tag{4.3}$$

$$\operatorname{sgn}[C_{l\_d}(x, y)] = \begin{cases} 1 & C_{l\_d}(x, y) \geq 0 \\ -1 & C_{l\_d}(x, y) \leq 0 \end{cases} \tag{4.4}$$

The right selection of sub-band coefficients plays a key role in the computation of well discriminative feature representation. It has been observed through extensive experimentation that the sub-bands acquired from vertically oriented CT directional filters contain enough discriminative information especially in the case of human silhouette that it can be accurately differentiated from vehicles. Therefore, these sub-bands have been exploited only to extract shape features. This reduces the number of CT sub-bands from 28 to 14 and the size of feature descriptor is eventually reduced.

After the computation of CT sub-bands, the next step is the extraction of features to represent the pattern. This has been done by computing LBPs on CT sub-bands.

Following [190], the value of the threshold T is selected to be 1% of the pixel value range that is 0.01 for data that lies between 0 and 1. The CT sub-band is divided into cells with a location grid in order to incorporate spatial information. A Cartesian grid of 4×4 cells is applied to each sub-band. For each cell, a CS-LBP histogram is computed. In order to avoid boundary effects in which the descriptor abruptly changes as feature shifts from one histogram bin to another, the bilinear interpolation is used to distribute the weight of each feature into adjacent histogram bins.

The final descriptor for each sub-band coefficient is computed by concatenating the CS-LBP histograms of all the cells. The resulted feature vector from each CT sub-band has a dimensionality of 256 ($16 \times 4 \times 4$). Since, 14 CT sub-bands have been used here, the resulting feature vector that is formed after concatenating the CS-LBP from all sub-bands has a dimensionality of 3584 (14×256). Following [190], the feature vector is normalized to unit length to reduce the influence of very large descriptor elements. Figure 4.3 comprehensively illustrates the computation process of CCS-LBP descriptors.

4.3.3 Classification

The CCS-LBP feature vectors of the training images are extracted and labelled accordingly. A non-linear SVM in one-vs-all manner is trained using the feature vectors of the training data using Radial Basis Function (RBF) kernel. The best combination of Soft Margin parameter C and gamma γ has been selected by a grid search with exponentially growing sequences of C and γ. The selection of RBF kernel and non-linear SVM has been done empirically.

4.4 Detailed Annotation of Moving Objects

4.4.1 Humans

The annotation of detected human figure is carried out in few stages to extract various details. At the initial stage, three main parts (i.e. head, torso and lower limbs) of the body are separated. This is followed by a colour extraction stage where the dominant colours in each portion of the body are extracted to analyse the clothes. Finally, clothes of human figures are analysed for the presence of text. The details about colour extraction, text detection and text recognition approaches proposed are given in chapter 5 and chapter 7, respectively.

A simple strategy has been devised here to categorize various parts of the human body. The aim is to categorize three main parts (head, torso, lower limbs) of the human silhouette. It is assumed that a human figure is in standing position (i.e. foreground humans are those that are walking). The challenging part is the estimation of head region. For that, two novel criteria are used:

Criterion I: It is subjectively examined on a significantly large collection of images containing humans that the head region occupies the approximately upper 12.5% of the human body. This criterion works well in almost all the cases with minor errors associated with the change of pose and camera angle. However, overall the location of head region is successfully estimated.

Criterion II: In order to reduce the above mentioned error the vertical projection of the binary map of human figure is analysed. The proposed foreground extraction algorithm (in section 4.1) computes an accurate binary map of the moving objects. The vertical projection graph of that binary map is exploited here to find the head region. The vertical projection graph is computed and a smoothing operation is applied to enhance the dominant peaks and valleys in it. The smoothing of the data is performed using a 5-point moving average filter. Consider $x(k)$ as the k-th data point, then using an n-point moving average filter the smoothed value $x^{'}(k)$ is computed as:

$$x^{'}(k) = \frac{1}{n} \sum\nolimits_{i=k-n+1}^{k} x(i) \tag{4.5}$$

Next, the top 20% part of the vertical projection graph is searched for a valley and the minima point in that valley is taken as the point of separation between head region and rest of the body. The reason for using 20% of the silhouette to search for the head is derived from the finding of criterion-I. Since, the dominant valley is likely to be located somewhere around the upper 12.5% of the human silhouette, the threshold of 20% provides a safe margin to deal with exceptional cases where the vertical projection is affected due to wrong foreground estimation or due to irregular human pose.

The two proposed criteria are used in a hierarchical manner. Preference is given to the estimation of criterion-II; if it fails to locate the neck region, criterion-I is used to estimate the point of separation between head and torso. For example, in some cases as shown in figure 4.4, there is no valley in uppermost 20% vertical projection. These cases arise when a person's neck is fully covered with hair or clothes. In such cases, the criterion-I has to be used to separate the head region in that human figure.

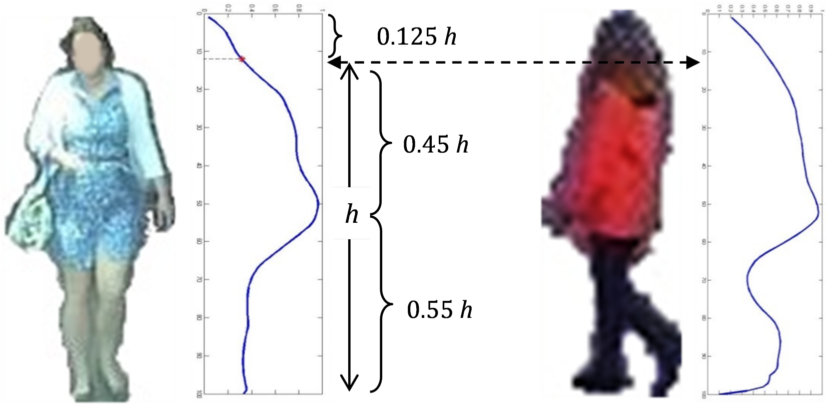


Figure 4.4: Detection of separation point between head region and torso in human figures along with the illustration of a scenario where no valley is detected in the uppermost 20% of the human body and the separation between head and torso is estimated using criterion-I.

Once the head region is estimated, it is easy to decide on the torso and lower limb regions. The remaining body is divided into two parts, the upper 45% of the remaining body is selected as the torso and the lowermost 55% as the lower limbs. This criterion is illustrated in figure 4.5 and is empirically tested on a large number of human images. The result indicate that it works exceptionally well in estimating the main regions of the human figure in standing (or walking) positions.

4.4.2 Vehicles

The vehicle object is annotated on the basis of license plate, colour and make & model information. It is assumed in this research that the front view of the vehicle is captured by the camera. The proposed annotation scheme relies on the detection of license plate regions therefore the presence of that information is important. Apart from that, the vehicle object comprises of various parts i.e. tires, windscreen, windows etc. that appear black and do not contain useful information that can be used for recognition and annotation of that vehicle.

The annotation of vehicle objects begins with the extraction of number plate region. The idea is to extract the RoI of the frontal view of the detected vehicle object and compute various attributes using the appearance information. A significant number of license plate detectors have been proposed in the literature [124, 123, 112]. This work follows the strategy proposed in [124] where the connected components are extracted and their geometric properties are analysed for the initial screening of candidate regions. In the next step, the shape features

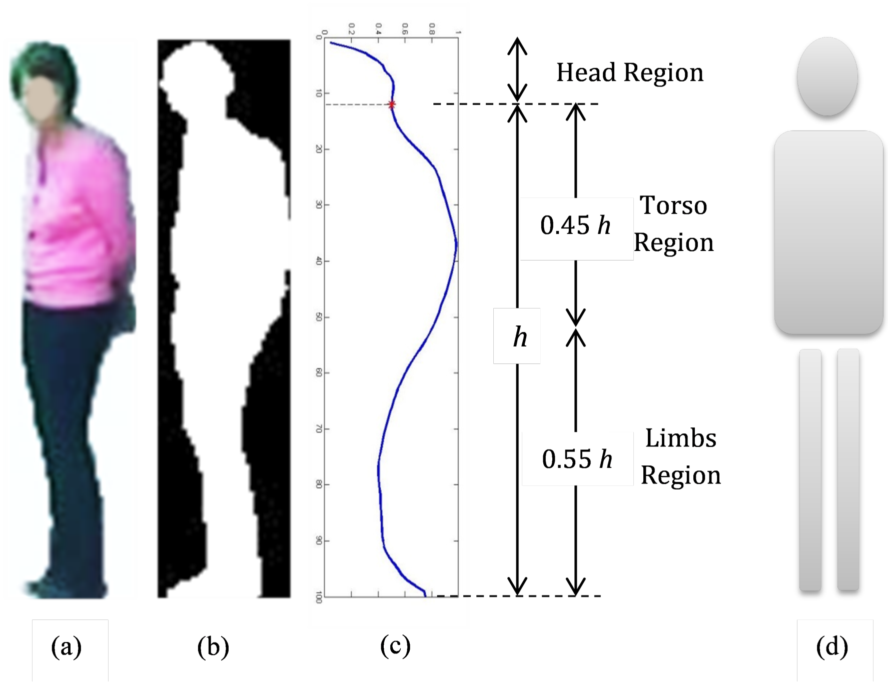


Figure 4.5: Categorization of main regions of a human figure. (a) Original Image. (b) Binary map of the estimated foreground. (c) Vertical Projection. (d) Estimated Silhouette

of the candidate regions are extracted and classified using a histogram matching technique. After the detection of license plate, the RoI that is extracted is used for the recognition of colour, make, model and type of the vehicle.

The license plate is detected and the location of the detected license plate is used as a reference to extract RoIs for computing the other appearance information. The RoI for the recognition of vehicle's colour uses the width of the license plate. A window is selected just above the licence plate region to enclose mainly the front hood (bonnet) of the car as this area contains mostly the pixels with actual colour of the car's body. Let w be the width of detected license plate, an region of width $3w$ and height w is selected as RoI exactly above the license plate as shown in figure 4.6. The pixels in the RoI are used for colour recognition of the vehicle.

4.5 Experiments and Results

The detailed experimental analysis of the novel feature descriptor and the proposed improvements in the foreground extraction are presented in this section. The benchmark datasets have been used for experiments in each section. To measure the accuracy of the methods the following standard metrics have been employed:

- Recall: Rec = TP / (TP + FN)

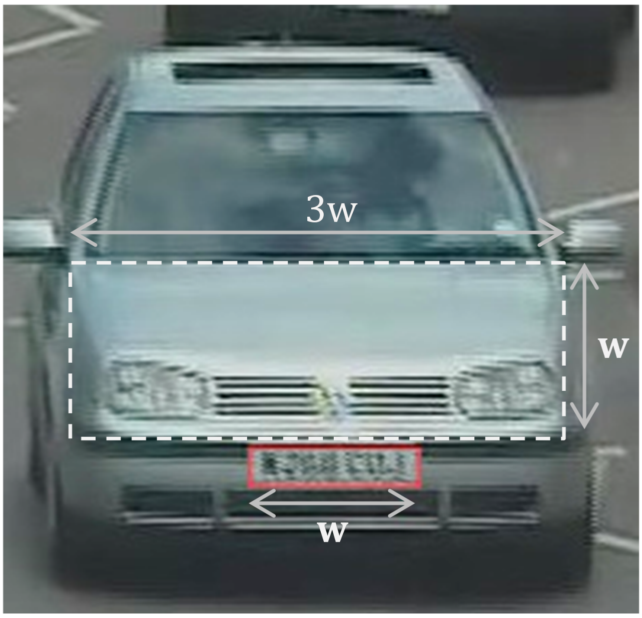


Figure 4.6: Selection of car region for colour recognition.

- Precision : Prec= TP / (TP + FP)
- Specificity: Spec= TN / (TN + FP)
- False Positive Rate: FPR= FP / (FP + TN)
- False Negative Rate: FNR= FN / (TP + FN)
- F-Measure : F= (2 * Prec* Rec) / (Prec + Rec) which is the weighted harmonic mean of Precision and Recall, so it can be regarded as an overall accuracy measure.
- False Positive Per Window (FPPW): False Alarms/ Total Testing Negative Examples

where, TP: True Positive FP: False Positive FN: False Negative TN: True Negative.

The details about the datasets and the experimental setup are given in following sections.

4.5.1 Foreground Segmentation

The performance of the proposed foreground extraction framework is tested on two standard datasets: CAVIAR [43] and ChangeDetection.net [51]. Apart from that, a comprehensive dataset has been collected to aid the research work presented in this chapter. The dataset is compiled from video clips of real CCTV feeds. These videos were acquired from a local CCTV surveillance centre. The video clips are carefully selected and manually labelled with the ground truth information for vehicles and humans. The dataset is named as DIRG (Digital Imaging Research Group) dataset.

In order to compare the performance of the proposed foreground extraction technique, the method presented in [69] is used as the baseline technique. The implementation of the selected baseline technique is publicly available in an open source library named OpenCV [13]. According to [69] the initialization of the background requires the provision of 4 parameters; the number of Gaussians in the mixture K, the portion of distributions accounted for by the background T, the initial standard deviation σ and the learning rate a. The values of these parameters in this study were empirically selected as follows: $K = 4$, $T = 0.55$, σ=30, $\alpha = 0.004$ for the processing of coloured frame and $K = 2$, $T = 0.45$, $\sigma = 40$, $\alpha = 0.03$ for the processing of gradient frame. The learning rate values for the two background models were set in such a way that if a foreground object becomes stationary (i.e. the moving vehicle or the walking persons stops) then it is absorbed by the background of the two models at the same time. The foreground segmentation process is performed on frames that have RGB components ranging between 0 and 255.

Table 4.1 presents various performance metrics and compares the proposed technique with the baseline method on CAVIAR dataset. Table 4.2 compares the performance on ChangeDetection.net dataset. The frames selected from the ChangeDetection.net dataset belong to cases "pedestrians" and "backdoor". It is evident from the results that the proposed method improves all metrics including the F-measure that depicts the amount of overall improvement. The decrease in the False Positive Rate (FPR) and False Negative Rate (FNR) shows that the proposed method successfully eliminates the shadow region while enclosing the incomplete foreground in the detected contour. The high value of recall and specificity indicates the fact that the True Positive Rate (TPR) and True Negative Rate (TNR) is improved without causing a degradation impact on the overall precision.

Table 4.1: Performance comparison of the baseline technique and the proposed method on CAVIAR dataset.

	Method	
	Kaewtrakulpong and Bowden [69]	Proposed
Recall	0.8391	0.9658
Specificity	0.9798	0.9812
Precision (PPV)	0.7429	0.8110
FPR	0.0202	0.0188
FNR	0.011	0.0029
F-measure	0.7881	0.8817

Table 4.2: Performance comparison of the baseline technique and the proposed method on ChangeDetection.net dataset.

		Recall	Specificity	Precision	FPR	FNR	F-Measure
Kaewtrakulpong [69]	Pedestrian	0.95	09997	09594	0.0003	0.0004	0.9546
	Backdoor	0.9073	0.9864	0.5317	0.0136	0.0016	0.6705
	Average	0.9287	0.9931	0.7456	0.0070	0.001	0.8126
Proposed Method	Pedestrian	0.9647	0.9995	0.9435	0.0005	0.0003	0.9540
	Backdoor	0.9126	0.9994	0.9659	0.0006	0.0015	0.9385
	Average	0.9387	0.9995	0.9547	0.0005	0.0009	0.9463

A visual comparison of the proposed method and the baseline technique (in figure 4.7) reveals that the baseline method is prone to incomplete and inaccurate foreground segmentation and shadow detection. In the first two examples it is noticeable that the gradient based method is capable of removing such shadow regions that do not form a distinctive outline. The example from the “backdoor” sequence shows that the proposed method successfully eliminates the shadows and false foreground regions that occur due to change in illumination. The pixels in the binary map of the foreground regions appear well connected and the gaps are properly filled at the same time. As the main objective here is to acquire the foreground segmentation without any breakage and gaps, the probability of false foreground detections is also increased. Another shortcoming of the proposed system is shown in the second example where the lack of gradient information (i.e. similarity of the foreground-background) results in incomplete foreground segmentation.

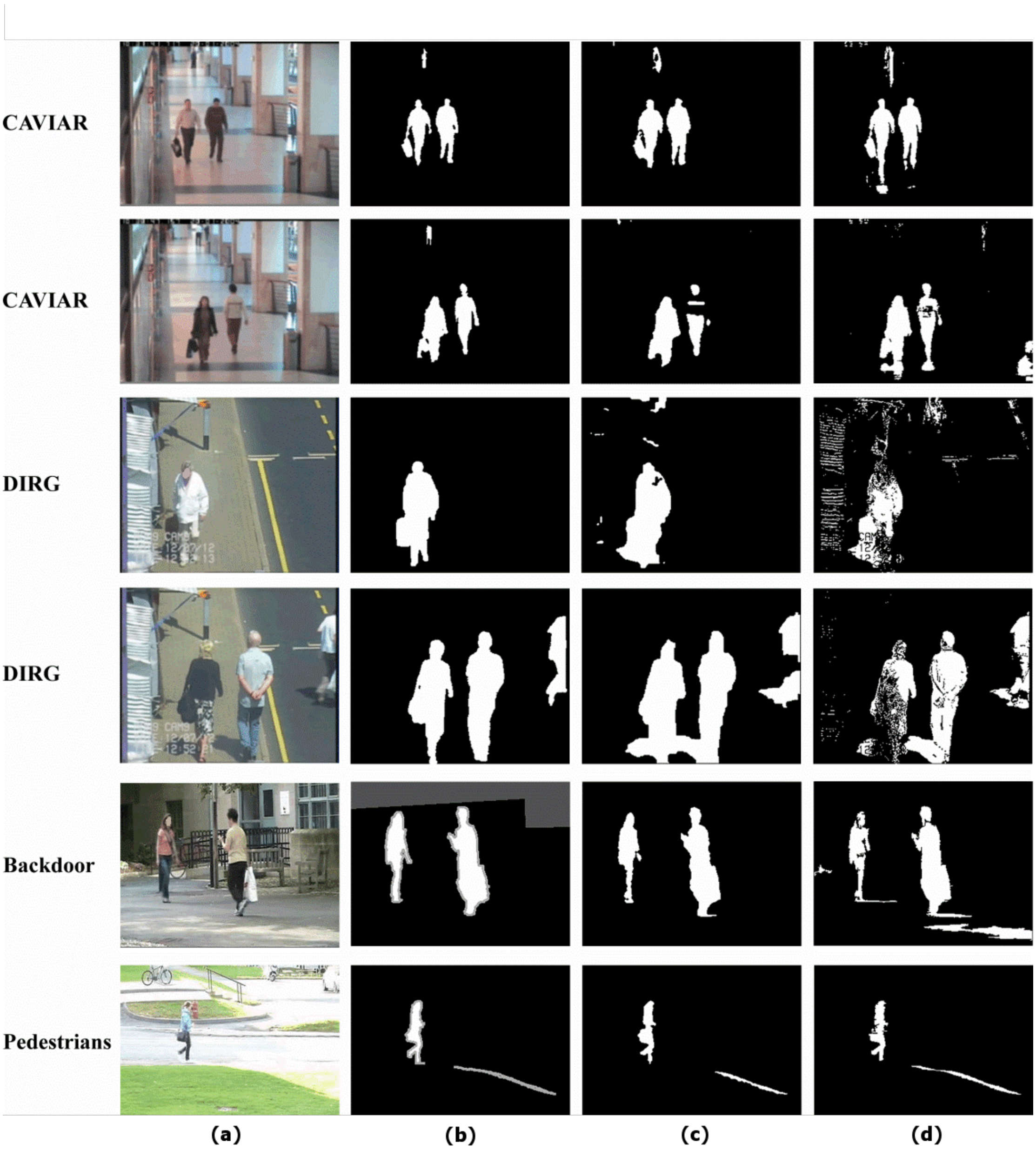


Figure 4.7: Visual comparison of the proposed method and the baseline technique for accurate foreground segmentation. (a) Input video frames. (b) Ground truth (c) Output using the proposed method (d) Output using the baseline method.

4.5.2 Recognition of Humans and Vehicles

In order to test the performance of the proposed CCS-LBP feature descriptor, two sets of experiments have been performed. In the first set of experiments the performance of the proposed descriptor is analysed with different input parameters i.e. no. of CT sub-bands, image size etc. to find the best set of parameters for the optimized performance.

The second set of experiments analyses the performance of the proposed feature descriptor on two challenging datasets: DIRG dataset and INRIA person's dataset [32]. Since, the DIRG datasets comprise of video clips, the experiments have been executed in conjunction with the proposed foreground extraction method. The output of the foreground extraction module from each video frame is fed into the object identification framework.

The DIRG dataset comprises of 3000 images (1000 human images, 1000 vehicle images and 1000 noise images). The test set comprise of 7000 images (2000 human image, 2000 vehicle images and 2000 noise images). These images have been arbitrarily extracted from the video frames.

On the other hand, the INRIA dataset comprises of human and non-human images. Therefore, the human object detection has been performed in a sliding window based detection setup. The windows of various sizes are extracted from the test image and fed into the proposed object identification setup where they are classified as humans or non-humans.

In order to compare the performance of the proposed framework the HOG-SVM based human detection technique[32] is used as baseline.

Optimized Parameters for CCS-LBP Feature Representation: The effect of different choices of orientations and decomposition levels of the CT coefficients is studied. It is examined experimentally that the correct choice of sub-bands is essential for achieving superior performance in the proposed object identification framework. It is mentioned in [38] that a l-level directional filter tree decomposition can be viewed as a 2^l parallel channel filter bank with equivalent filters and sampling matrices. This corresponds to separable sampling matrices grouped into two ranges of the l-level directional sub-bands in each pyramid decomposition level.

Accordingly the two sets correspond to 0 to 2^{l-1} and 2^{l-1} to 2^l sub-bands in each pyramid decomposition level. These two sets convey the horizontally oriented and vertically oriented set of directions. In the context of this work, this translates that at each pyramid level the second half of the corresponding directional sub-bands would convey the vertically oriented directions. The density of vertical edges is significantly high in a human figure; therefore the experimental emphasis is given to those CT coefficients that are achieved from vertically oriented filters of the filter bank. The vertically oriented edges have higher strength in these CT coefficients while the edges lying in other directions are completely suppressed.

Apart from the selection of CT sub-bands, the selection of input image size is also important for the improvement of identification results. The FPR vs TPR curves in figure 4.8 show that the best identification performance is achieved when the second half of sub-bands from 3-level CT decomposition is selected using an image size of 256 × 256.

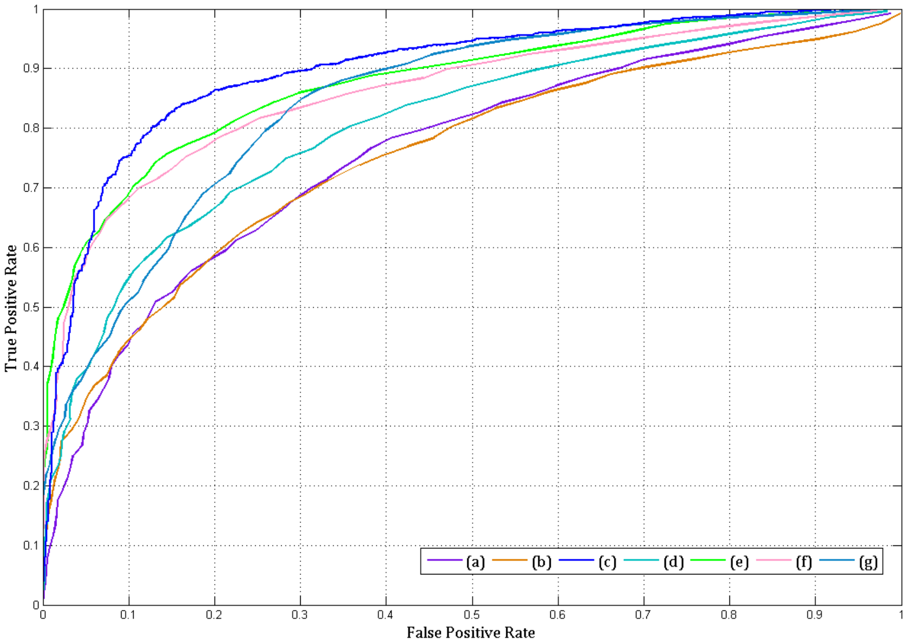


Figure 4.8: The performance comparison of various band selections and sizes of input image. (a) 4 sub-bands at 1st and 3rd level of CT decomposition and 128 × 128 input image. (b) 4 sub-bands at 1st and 3rd CT decomposition level and 256 × 256 input image. (c) 14 sub-bands at 3 CT decomposition levels and 256 × 256 input image. (d) 14 sub-bands at 3 CT decomposition levels and 128 × 128 input image. (e) 6 sub-bands at 2 CT decomposition levels and 256 × 256 input image. (f) Rotation Invariant Uniform LBP on 14 sub-bands at 3 CT decomposition levels and 256 × 256 input image. (f) 14 sub-bands at 3 CT decomposition levels and 512 × 512 input image.

Recognition Results with CCS-LBP Features: The Receiver Operating Characteristic (ROC) curves on DIRG dataset in figures 4.9 show that the proposed CCSLBP-SVM technique outperforms the baseline (HOG-SVM) technique by a significant margin. As shown in Table 4.3, the proposed method has achieved substantially high values of Precision, Recall, Specificity and F-measure clearly indicating the superiority of the proposed methodology. The FPR and FNR values on the other hand have decreased that shows the reduced number of false negatives and false alarms. High values of Recall and Specificity indicate that TPR and TNR have improved without affecting the Precision.

As mentioned before, the experiments on INRIA dataset have been performed in a sliding window based detection setup. The same testing setup as in [32]

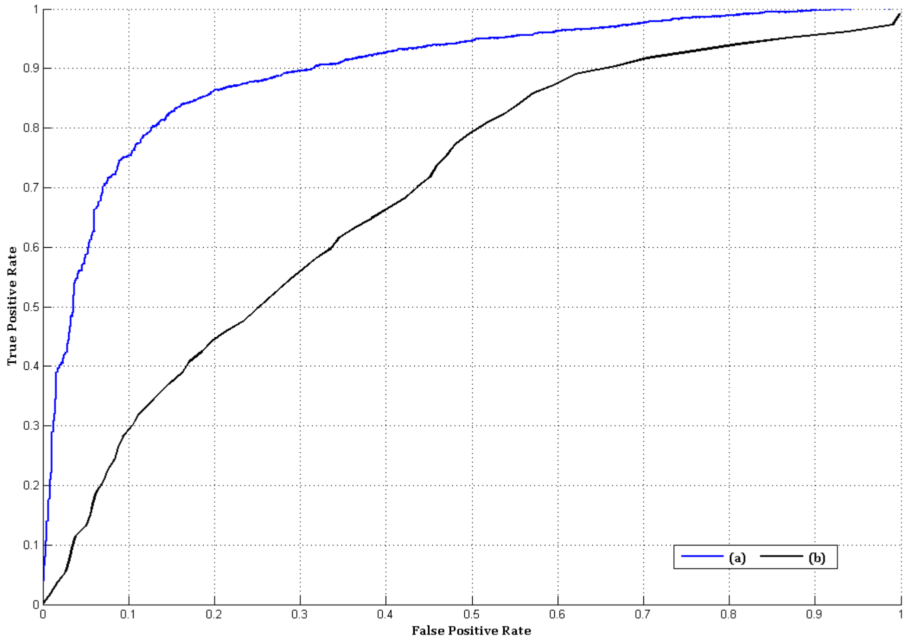


Figure 4.9: ROC Curves showing the performance comparison between the proposed object identification technique (CCSLBP-SVM) and baseline technique (HOG-SVM). (a) CCSLBP-SVM. (b)HOG-SVM.

Table 4.3: Performance metrics for the comparison of the proposed method and the baseline (HOG-SVM) technique on DIRG dataset.

	Method	
	HOG-SVM (Baseline)	CCSLBP-SVM (Proposed)
Sensitivity	0.433	0.784
Specificity	0.810	0.818
Precision (PPV)	0.942	0.970
F-measure	0.593	0.872
Accuracy	0.479	0.788
FPR	0.190	0.182
FNR	0.567	0.216

and [72] is adopted here. The windows of 128 × 64 pixels were extracted at three different scales (1, 1.2 and 0.6 times) of the original image. The sliding window is moved by a stride of 4 pixels in first two scales (1 and 1.2) and by 2 pixels in the third scale (0.6) of the image. The proposed feature descriptor is extracted on each window and classified using the trained SVM classifier. Finally the windows that were classified as humans across three scales of the image and had significant overlap were merged together indicating a single human subject. Table 4.4 compares the Miss Rate at the False Positive per Window (FPPW) of 10^{-4} of the proposed technique with other methods. The performance comparison

Table 4.4: Miss rate of the proposed methods and other techniques at the FPPW of 10^{-4}.

Method	Dalal and Triggs [32]	Kim et al. [72]	Proposed
Miss Rate	0.172	0.162	0.159

shows the strength of the proposed feature descriptor, it superseded the other techniques.

4.6 Summary and Conclusion

A framework for the detection and recognition of moving objects in CCTV surveillance videos is presented in this chapter. The main contributions of this research work involves an improved foreground extraction technique, a novel feature descriptor for the classification of humans and vehicles and a technique for the classification of human body parts into head, torso and leg regions. The proposed techniques have been tested on benchmark datasets and compared with existing state of the art methods.

The improved foreground segmentation technique accurately extracts the foreground object and its contour in the presence of cluttered background. It also fills gaps in the binary map of the foreground regions. The technique proves to be robust and eliminates the surrounding noise that causes inaccuracies in extracted foreground regions and increase processing overhead. The experimental results show that the improved technique achieved better performance as compared to baseline technique [69].

Apart from foreground extraction, the proposed CCS-LBP feature descriptor provides a discriminative representation of the objects. It exploits the fact that human silhouette contains a large number of contour segments and use the pattern of these contour segments to classify the objects into humans and vehicles. The experimental results showed the superiority of the proposed technique over the baseline method.

Chapter 5

Object Colour Correction and Recognition for Annotation and Retrieval

5.1 Introduction

The colour extraction of the objects in video frames involves various challenges. The first and foremost challenge is the correction of true object colours. The colours of objects especially in surveillance videos are significantly affected by the poor ambient illumination and inaccuracies of the calibration of camera sensors. For example, if a video is recorded during a bright sunny day, the true colours of the objects are influenced by the colour of sunlight, that may differ from being “white” at certain times of the day and under some extreme environmental conditions. Further during other times the surrounding ambient illuminant of an object influences the true colours and poses a challenge for any colour descriptor to capture the true colours present in the object of interest. The recently proposed approaches for colour recognition [23, 81, 15, 177] involve pixel clustering based on histograms or feature space analysis to grow the tiny regions of similar colours into bigger dominant regions. However, the problem of accurate colour recognition still requires significant improvement to achieve reliable accuracy especially in the case of medium to low quality surveillance videos where most of the content is dominated by noise.

To solve the above issue this chapter propose a framework that combines the advantages of various approaches [15, 145, 15] into a novel framework to address the following tasks:

1. Correction and enhancement of colour in video frames.

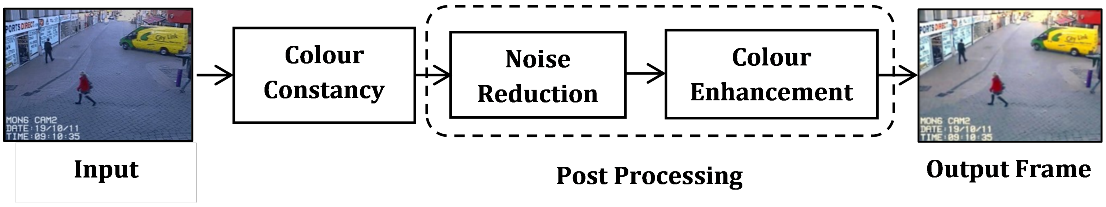


Figure 5.1: Block diagram of the colour correction and enhancement pipeline.

2. Recognition of dominant colours in the objects of interest.

The work presented in this chapter interpolates a portion of work presented in [48] and [49]. The proposed colour correction procedure exploits colour constancy algorithm to estimate the illuminant with in the image. The estimated values are used to cancel the effect of that illuminant so that the true colours of the objects become visible. The colour constancy procedure reduces the effect of surrounding illuminant but the true colours of the objects require a post processing enhancement step in order to be recognized correctly. The post processing procedure increases spatial difference between the pixels of various colours and suppresses the noisy pixels. The presence of noise negatively impacts the accuracy of dominant colour extraction.

An adaptive framework is therefore devised to perform the correction and enhancement of true colours of objects within video frames. The proposed framework takes advantage of an existing colour constancy technique and analyses each frame on the basis of its luminance and chrominance information to estimate the suitable parameters for the improvement of true object colours. In the next step, the colour recognition pipeline transforms the pixels from RGB to HSV space and quantizes the HSV space into a finite number of bins. The pixels are assigned to respective bins based on their HSV values and the dominant colour descriptor is computed to recognize the colour of the objects. The proposed framework and the experimental findings are explained in detail in the following sections.

5.2 Correction and Enhancement of Colours in Video Frames

The block diagram of the proposed colour correction and enhancement pipeline is shown in figure 5.1. Each stage is explained in detail as a separate sub-section.

5.2.1 Correction of Colour

Colour constancy is a key step in order to reduce the effect of illumination and surrounding reflection. It is impossible for a colour recognition system to perform well without colour constancy.

In [11], the authors empirically analysed and compared a number of well-known and widely used colour constancy algorithms that are based on colour image statistics. They showed by experimentation on benchmark datasets of synthetic and real images that there is no specific algorithm that works better than other algorithms in all scenarios. This indicates that all the existing techniques perform with reasonable accuracy and any of these techniques may be used to perform colour constancy in surveillance videos. However, the complexity and computational performance of different colour constancy algorithms is considerably different (see section 5.4.1). This leads the selection criterion to be the processing time and ease of implementation. Based on this criterion, the Gray World (GW) [16] algorithm became the preferred choice in this work for the estimation of surrounding illuminant within a video frame.

The Gray World algorithm is based on the assumption that the average reflectance in a scene is achromatic. The implementation framework used in this work was proposed by Van de Weijer et al. [155]. They unified various illuminant estimation algorithms using the instantiations of the following equation:

$$\int\int \left(|\nabla^n \rho_\sigma(x,y)|^p dxdy\right)^{\frac{1}{p}} = kI \tag{5.1}$$

where n is the order of derivative, p is the Minkowski norm, $\rho_\sigma(x,y) = \rho(x,y) * G_\sigma(x,y)$ is the convolution of image with a Gaussian filter $G_\sigma(x,y)$ with scale parameter σ and k is a constant that is chosen in such a way that the illuminant colour I has unit length. Gray World can be generated with the setting $(n,p,\sigma) = (0,1,0)$ in equation 5.1. The estimated illuminant I is used for the correction of colours in the input video frame. Following [155], the effect of surrounding illuminant is cancelled to acquire the true colours of the objects.

The colour constancy algorithm significantly reduces the effect of illuminant, however, it does not guarantee the recovery of true object colours especially if the input image is slightly over or under exposed. This instigates the need for a post processing stage that reduces noise and enhances the strength of true colours of objects within the video frames. The key steps involved in the post processing stage are explained in following section.

5.2.2 Post Processing Pipeline

Noise Reduction: A major issue with the processing of surveillance videos is the strong effect of noise. The noisy pixels in video frames occur due to various factors. These factors include compression, low resolution and deficiencies in camera sensors. Typically surveillance videos are recorded at a low resolution and the

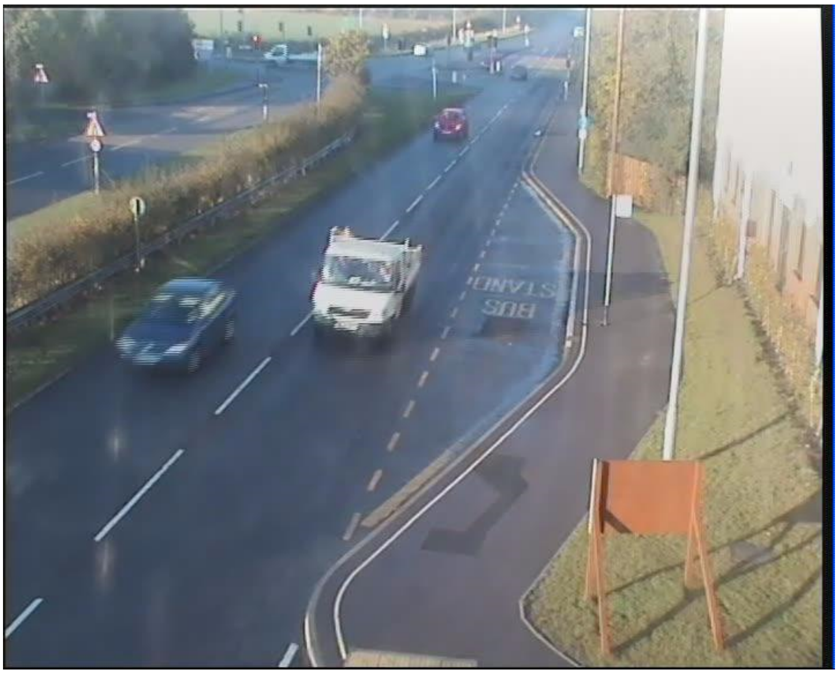


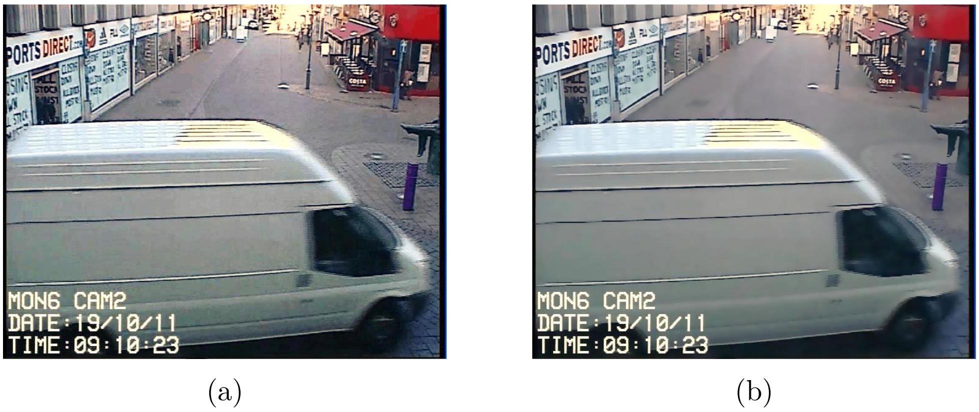
Figure 5.2: A CCTV frame captured at low resolution and affected by noise. Notice the lines in the moving objects depicting the effect of interlacing.

highest possible degree of compression is applied to reduce the size of video data. The compression procedures involve interlacing where the information present in half of the frame is virtually removed. Figure 5.2 shows an example of such video frame. Notice the effect of interlacing in the form of horizontal lines in and around the edges of vehicles. All these factors result in the loss of details. This makes the processing and extraction of any appearance information from these frames a significantly challenging job. Therefore, the reduction of noise and recovery of lost information is an important step to efficiently extract and recover the visual details present in the video frames.

The reduction of noise is often a challenging task since the common noise reduction filter such as an averaging filter although successfully merges the noise pixels with their neighbouring pixels but fails to preserve edges. This result in a phenomenon called colour bleeding, where the effect of the colour of one region propagates to the adjacent region causing a contamination to the true colour of the adjacent region and vice versa. This demands a noise reduction strategy that preserves edges. Therefore, Bilateral Filtering (BF) [152] based noise reduction has been adopted here.

Consider I being the image achieved after the colour constancy stage. The smoothed image I_{sm} obtained using a BF according to [152] is given as:

$$I_{sm}(x) = k^{-1} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} I(\vartheta) g(\vartheta, x) p(I(\vartheta), I(x)) d\vartheta \tag{5.2}$$



(a) (b)

Figure 5.3: The output of noise reduction step. (a) Input Video frame (b) Output after Bilateral Filtering.

with normalization,

$$k(x) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(\vartheta, x)p(I(\vartheta), I(x))d\vartheta \tag{5.3}$$

where $g(\vartheta, x)$ measures the geometric closeness between the neighbourhood centre x and the close-by point ϑ, $p(I(\vartheta), I(x))$ measures the photometric similarity between the pixel at the neighbourhood centre x and that of a nearby point ϑ. Figure 5.3 shows a video frame before and after noise reduction. It is noticeable that the strength of the dark grey noise pixels present in the white vehicle (figure 5.3a) is reduced and the pixels are merged with the surrounding white pixels (figure 5.3b).

Enhancement of True Colours: The goal of this stage is to improve the lighting and colour information within the video frames so that the true colours of the objects are easily recognizable using colour space quantization and dominant colour extraction. The video frame acquired after the noise reduction stage is transformed to the HSV colour space. The HSV representation is more relevant to human perception of colours unlike the RGB space, which is non-absolute and was mainly designed for display devices such as monitors etc. The details about the selection of bins and angular cut off values are given in section 5.3.1.

The histogram of Saturation S and Value V components of the frame are analysed. This is done by modelling the data as a random variable. The probability density function is computed using non-parametric kernel density estimation. Let $(x_1, x_2, ..., x_n)$ be the histogram bins of random variables (V and S in this case) with an unknown density p. The estimated shape of this function p using kernel density estimator proposed in [133] is given as:

$$p_h(x) = \frac{1}{nh} \sum\nolimits_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{5.4}$$

Here, $K(.)$ is the kernel, n is the number of bins (100 bins ranging from 0 to 1) in the V histogram, h is the bandwidth and K_h is the scaled kernel defined as: $K_h(x) = (1/h)K(x/h)$. The density estimation of S is computed using the same procedure. In order to perform density estimation having a normal distribution kernel function $K(.)$, the smoothing factor h is computed as suggested by Silverman [141]. Hence,

$$h = \hat{\sigma} C_{\upsilon}(k) n^{-1/(2\upsilon+1)} \tag{5.5}$$

where $\hat{\sigma}$ is the sample standard deviation, υ is the order of the kernel, and $C_{\upsilon}(k)$ is the constant for the chosen distribution, which for a Gaussian kernel of size $\upsilon = 2$ is equal to 1.06. Therefore it can be written as:

$$h = 1.06 \hat{\sigma} n^{-1/5} \tag{5.6}$$

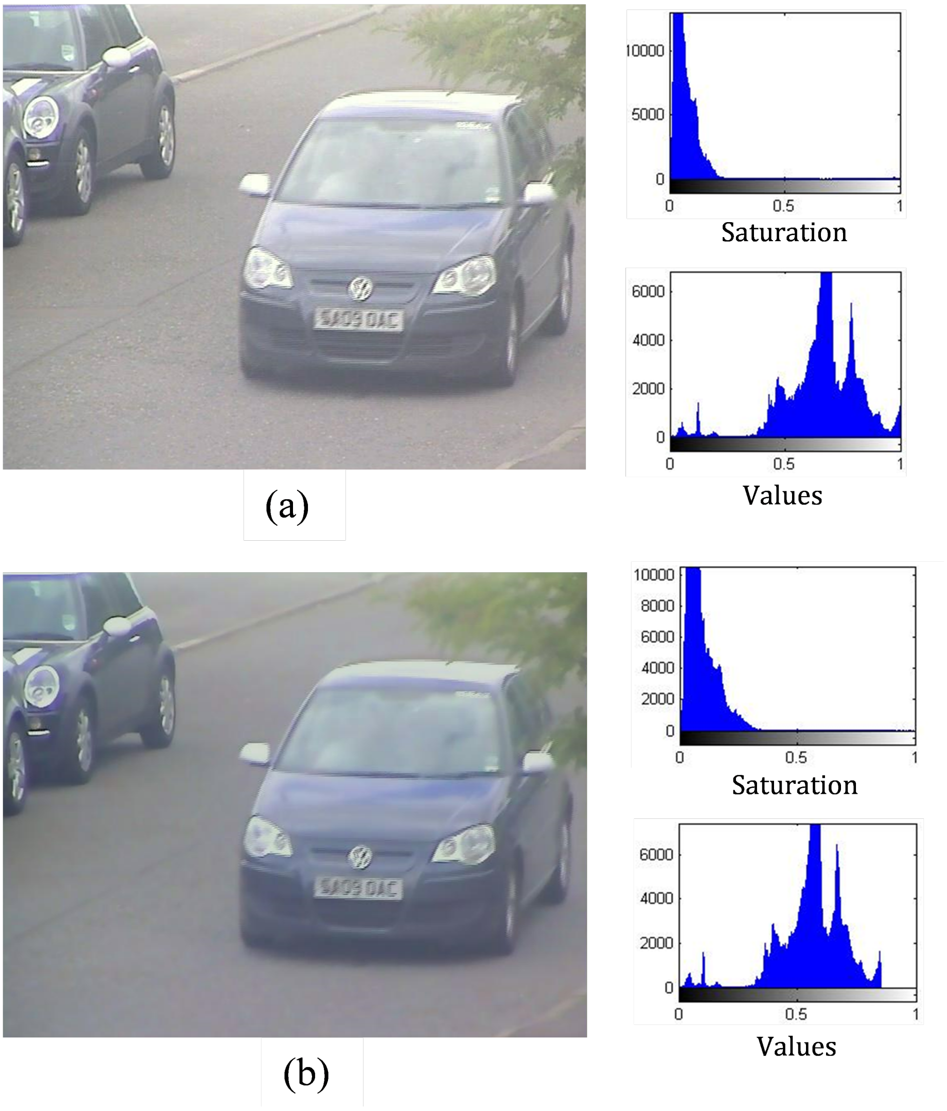
The samples are separated into groups as indicated by the local minima of $p_h(x)$ and each sub-sample has its mean at the local maxima of $p_h(x)$.

The estimated Probability Density Functions (PDFs) $p(s)$ and $p(v)$ of S and V histogram respectively provide information about the distribution pattern of chrominance and luminance in the image respectively. This information is used to compute suitable parameters to enhance the light and colour in the image. A set of rules is devised that exploits the mean and standard deviation of the most significant peak in the estimated densities to evaluate the quantity of correction required within the image for accurate colour extraction. These rules have been established after excessive data mining of real video frames and images.

Let, significant peak in $p(v)$ and $p(s)$ has the mean m_v and m_s respectively. Then:

Case I: If $m_v > 0.6$ and $m_s < 0.15$, the probability for the image under consideration to be over exposed is high. Therefore, it requires to be transferred to the lower value in order to reduce the effect of over exposure. This is corrected by shifting the density $p(v)$ to a new mean, $m_v^{'} = \alpha$. Apart from modifying m_v, the density $p(s)$ is also shifted to a modified mean $m_s^{'} = \beta$.

Case II: If $m_v < 0.6$ and $m_s < 0.15$, the probability for the image under consideration to be an underexposed image is high. Therefore, it requires an increase in the strength of higher values of the V component in order to attain a better separation among different colours. This is achieved by shifting the mean



][t]

Figure 5.4: Colour enhancement of an overexposed image. (a) Original image and the corresponding S and V histograms. (b) The enhanced image along with the corresponding S and V histograms after the application of the proposed colour enhancement procedure.

of the distribution $p(v)$ to $m_v^{'} = \alpha$ while the mean for the distribution $p(s)$ to $m_s^{'} = \beta$.

The empirically estimated values $\alpha = 0.58$ and $\beta = 0.16$ works best in case of surveillance videos. The estimated thresholds are applied to V and S components of the video frame in a non-linear manner expressed in equation 5.7 and 5.8 respectively. The reason for non-linear enhancement is to increase the separation among the colour pixels in adjacent bins. This is illustrated in 5.4 and figure 5.5 where the proposed colour enhancement mechanism is applied to an underexposed and overexposed video frame respectively, along with the original and modified S and the V histogram of each video frame.

Let the V component of the image after colour constancy be $V_{in}(\forall pixels)$. The

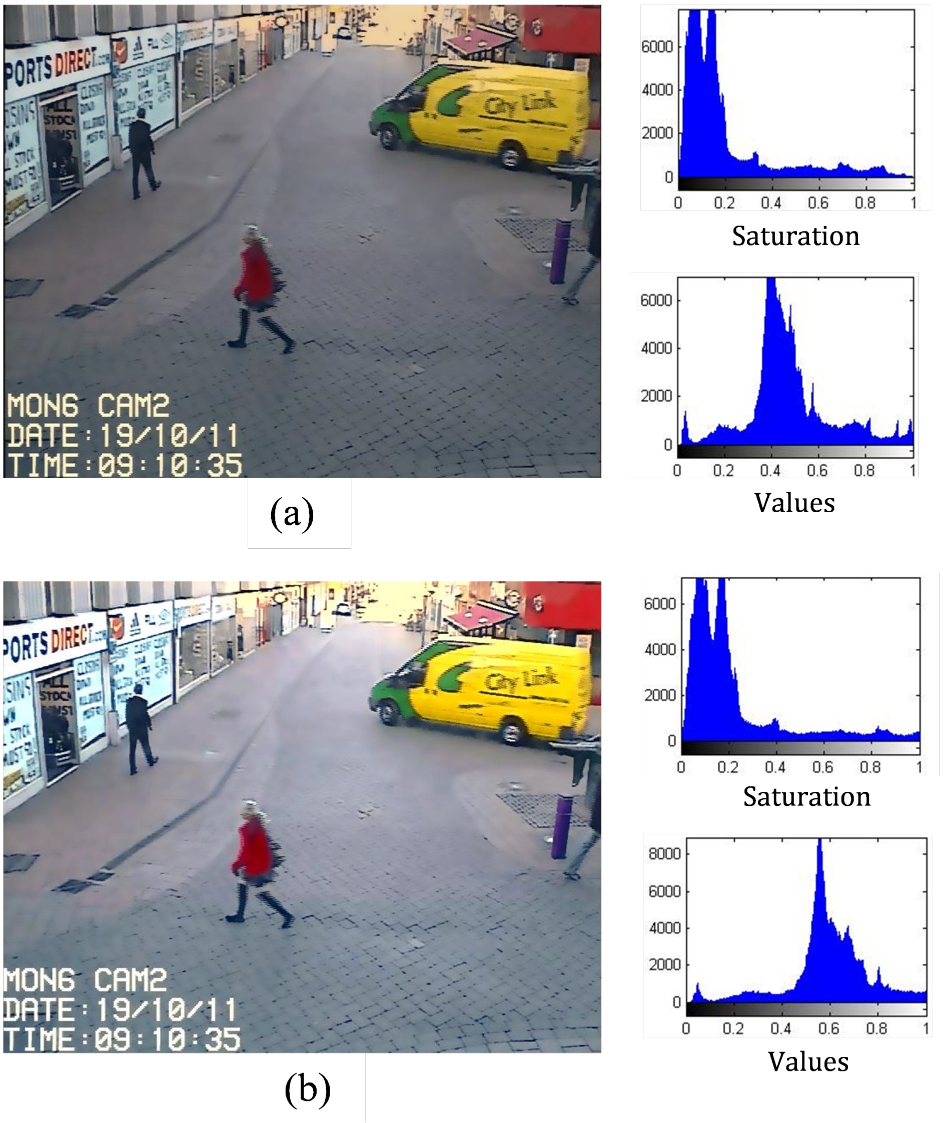


Figure 5.5: Colour enhancement of an underexposed image. (a) Original image and the corresponding S and V histograms. (b) The enhanced image along with the corresponding S and V histograms after the application of the proposed colour enhancement procedure.

enhanced V component V_{out} obtained using the adaptive framework is given as:

$$V_{out}(x, y) = \delta[V_{in}(x, y)]^{(1-\Delta m_v)} \tag{5.7}$$

where, $\Delta m_v = m_v^{'} - m_v$ and δ is a constant. The value $\delta = 1$ is used in the experiments.

In the same way, let the S component of the image after colour constancy be $S_{in}(\forall pixels)$. The enhanced S component S_{out} obtained using the adaptive framework is given as:

$$S_{out}(x, y) = S_{in}(x, y)\,(1 + |\Delta m_s|) \tag{5.8}$$

where, $\Delta m_s = m_s^{'} - m_s$.

The image acquired after the proposed post processing stage depicts better strength of the true colours of objects along with the suppression of the surrounding illumination and noisy pixels. This aids the subsequent process of dominant colour extraction to be performed accurately.

5.3 Recognition of dominant colours in the objects of interest

5.3.1 HSV Quantization

The main focus of this research work is the extraction of three dominant colours from the objects of interest.The first step in the extraction of the dominant colour is the quantization of the colour space. The quantization process is important to reduce the total number of colours in the colour space to a finite value. If the colour space is not quantized then the process will run virtually for an infinite number of colours that may result in the execution of the colour extraction process up to an infinitely long time without achieving any considerable results. Another important thing to consider is the selection of quantization levels. The quantity of the selected number of quantization levels should neither be too small nor too large. If the number of selected quantization levels is high then the purpose of quantization will not be served as the system will have to process a large number of colours. On the other hand, if the number of quantization levels is too small then it may result in a confusion between adjacent colours, which may eventually cause a poor recognition performance.

Following [135], the HSV space is quantized into 72 colour levels (8 Hue, 3 Saturation and 3 Value). The angular cut off values in the hue separate 8 primary colours, whereas the various shades of these primary colours are separated in respective Saturation and Value bins. In this way, each pixel in an object is assigned to one of the 72 HSV bins and each bin is represented by the average of H, S and V values present in it.

5.3.2 Extraction of the Dominant Colour Descriptor

The objects of similar colour may appear different in the footage of multiple CCTV cameras due to the factors mentioned earlier. As, this work is addressing the annotation of CCTV videos that are recorded in significantly varying conditions, a reliable mechanism is essential to represent the dominant colours of the objects. Instead of relying on simple histogram based dominant colour extraction, the proposed framework uses a compact description of the representative colours in the objects of interest. Following [26], the MPEG-7 Dominant Colour Descriptor (DCD) is computed using the quantized colour histogram computed in section 5.3.1. The computation of DCD not only eases the extraction of dominant colour but allows the efficient indexing and retrieval of objects in large databases.

The DCD is defined as:

$$D_{DC} = \left\{ \left\{ C_i, P_i, R_i \right\}, s \right\}; (i = 1, 2, ..., N) \tag{5.9}$$

where, C_i is a 3D dominant colour vector containing the component values of the colour space (H,S,V in this case), P_i is the percentage of each dominant colour and R_i is the colour variance that describes the variation of the colour values of the pixels in a cluster around the corresponding representative colour. The spatial coherence s is a single number that represents the overall spatial homogeneity of the dominant colours in the image and N is the number of dominant colours. The focus of this work is to extract dominant colours of the object. Therefore the value $(N = 3)$ is selected here.

Following [135], the spatial coherence and the variance parameter is neglected. Let, χ_i expresses the proportion of each colour in the HSV colour histogram, where $i = 0, 1, 2..., 71$. Let, W_j expresses the percentage in descending order to χ_i , taking first N colors as the dominant colors, the non-dominant colors are no longer required, namely

$$\chi_i = \begin{cases} \chi_i & \chi_i = W_j \\ 0 & \chi_i \neq W_j \end{cases} \tag{5.10}$$

Normalizing the first N dominant colours corresponding percentage yields,

$$\chi = \{ {\chi'}_i, i = 0, 1, ..., 71 \} \tag{5.11}$$

where,

$$ {\chi'}_i = \frac{\chi_i}{\sum\limits_{j=0}^{N-1} W_j} \tag{5.12}$$

Consider D_{DC1} and D_{DC2} are the dominant colour descriptors of image 1 and image 2 respectively, the similarity measure $\nabla(D_{DC1}, D_{DC2})$ between two DCDs is computed using the following equation:

$$\nabla \left(D_{DC1}, D_{DC2} \right) = \sum\nolimits_{i=1}^{71} \min \left(\chi_{1i}, \chi_{2i} \right) \tag{5.13}$$

The value of similarity measure lies between 0 and 1. If the colour of two images is similar then the value of $\nabla(D_{DC1}, D_{DC2})$ is close to 1.

5.3.3 Extraction of Dominant Colours using Temporal Information

It is important to note that the required colour information related to moving objects is available in more than one successive frame that may serve as an aid to enhance the recognition performance. Therefore, the temporal information is exploited to enhance the accuracy and confidence of the colour extraction mechanism. The DCD of an object is computed by utilizing the presence of that object in multiple frames. The pixels associated with the said object are collected from all the frames in which the object appear. The histograms in subsequent frames are blended using the criteria presented in [15]. Consider h_i be the histogram of the subsequent frame. The cumulative histogram (H_i) for the total number of $(i+1)$ frames according to [15] is:

$$H_1 = h_1 \tag{5.14}$$

$$H_{i+1} = \frac{i}{i+1} H_i + \frac{1}{i+1} h_i \tag{5.15}$$

In this way, the number of pixels associated with a particular object increases and the noise pixels are suppressed by the high concentration of object pixels resulting in a reliable DCD computation. This also reduces the probability of wrong colour recognition due to inaccuracies of the colour histogram in a few frames. The histogram estimation may also be affected by the presence of occlusion or an inaccuracy in the foreground extraction algorithm.

A further step of processing is included here to compute the actual colour of an object. The actual colour means the one that a human will perceive for that object. For example, in the case of a car, the black colour dominates due to the usual colour of the tyres, wind screen etc. However from a human perspective the colour of the car is the one that dominates in the body region of the car. In order to reduce the effect of these factors, a weighted strategy is adopted here that uses the prior information regarding the object shape (i.e. human or vehicle) and takes into account the above mentioned factors.

A few colours are given higher weights than others while computing the DCD of those regions. For example, the weights for brown and black colours are set higher for head regions as it is highly likely that one of these two colours will be present in that region considering a person is walking towards or away from the camera. Similarly, in the case of vehicles the weight for brown colour is set to a small value as the brown colour is very rare among vehicles. The alteration in the weights is easily accomplished by changing the percentage parameter P of the

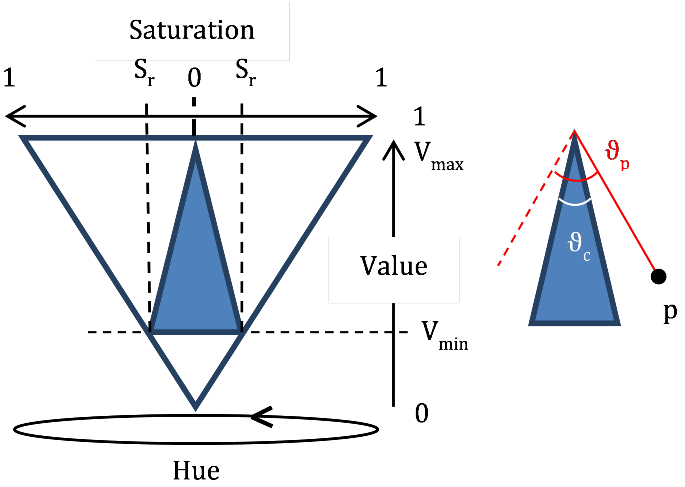


Figure 5.6: The dark region highlights the conic region of HSV plane where the pixels are more likely to belong to a white coloured object.

colours in the DCD before selecting the three dominant colours.

5.3.4 Weight Estimation for White Objects

The chance of error while dealing with white coloured objects is always high due to the fact that the white colour reflects its surrounding colours and may appear in the shades of blue, off white, gray or any other colour. For example, consider a scenario where a white vehicle is travelling on the highway during a sunny day, the white car reflects sunlight and most of the pixels present in the vehicle possess hue value closer to light yellow or light blue. In such cases, the dominant colour of the vehicle is estimated to be light blue or off white by any pixel clustering algorithm. In order to deal with this situation, a criterion is devised that estimates whether a particular object has a sufficiently high probability of being white.

The criterion is applied while computing the DCD of the object as explained in the previous section. The value of the average Saturation and the average Value is computed from the pixels of candidate object. If the average Value and average Saturation of the candidate object lies within a specific region in HSV then the chances are high that the object is white. The region is empirically estimated and is specified by the dark colour as shown in figure 5.6. In reality, this triangle is an inverted conic region in HSV space.

Let V_{avg} be the average of Value and S_{avg} be the average of Saturation of the pixels present in an object. In order to check whether the average value of the pixels of an object lies within the dark conic region of height ΔV, where $\Delta V = (V_{max} - V_{min})$, and radius S_r, the opening angle of the cone is compared against the opening angle formed by that point in the 3D space.

The opening angle ϑc (see figure 5.6) of the dark coloured cone can be computed using the following equation:

$$\vartheta_c = 2\tan^{-1}\left(\frac{S_r}{\Delta V}\right) \tag{5.16}$$

Consider a point p that lies in the HSV space with the coordinate $(S_{avg}, V_{avg}, H_{avg})$. The angle ϑ_p that the point will make at the vertex of the dark coloured cone can be computed using the following equation:

$$\vartheta_p = 2\tan^{-1}\left(\frac{S_{avg}}{V_{avg}}\right) \tag{5.17}$$

The point lies within or at the boundary of the dark coloured conic region if and only if $\vartheta_p \leq \vartheta_c$. This indicates that the chances for that object being a white coloured object is high. Therefore a high weight is assigned to the white colour while computing the dominant colour of that object.

5.4 Experiments and Results

The proposed framework is tested in a colour based object annotation and retrieval setup. The first set of experiments is performed on the videos collected from real CCTV footage acquired from Loughborough Town's CCTV control room. A set of 2 hours of video clips is separately extracted that contains 320 vehicles and 195 human figures. It was not possible to manually segment the foreground objects. Therefore the foreground extraction technique proposed in chapter 4 is applied to extract the binary map of the objects. Notice that the binary maps of a large number of objects did include some portion of the road and background due to the imperfections of the foreground segmentation algorithm. However, the temporal information helps to achieve good colour recognition accuracy. The objects were manually annotated for their colours based on subjective analysis. In human figures, the head, torso and leg regions were also labelled and the proposed colour recognition technique was applied on the labelled regions of each human figure.

Figures 5.7 and 5.8 demonstrate a few visual results for the extraction of dominant colours in humans and vehicles using the proposed colour recognition pipeline. In figure 5.8 the significant improvement in the dominant colour recognition is evident after the use of the proposed colour correction pipeline. The dominant colour recognition pipeline, prior the corrections, was unable to capture the true colours of the objects in any of the three extracted dominant colours. After applying the proposed colour correction pipeline, the true dominant colours of both objects have been successfully captured in the three most dominant colours.

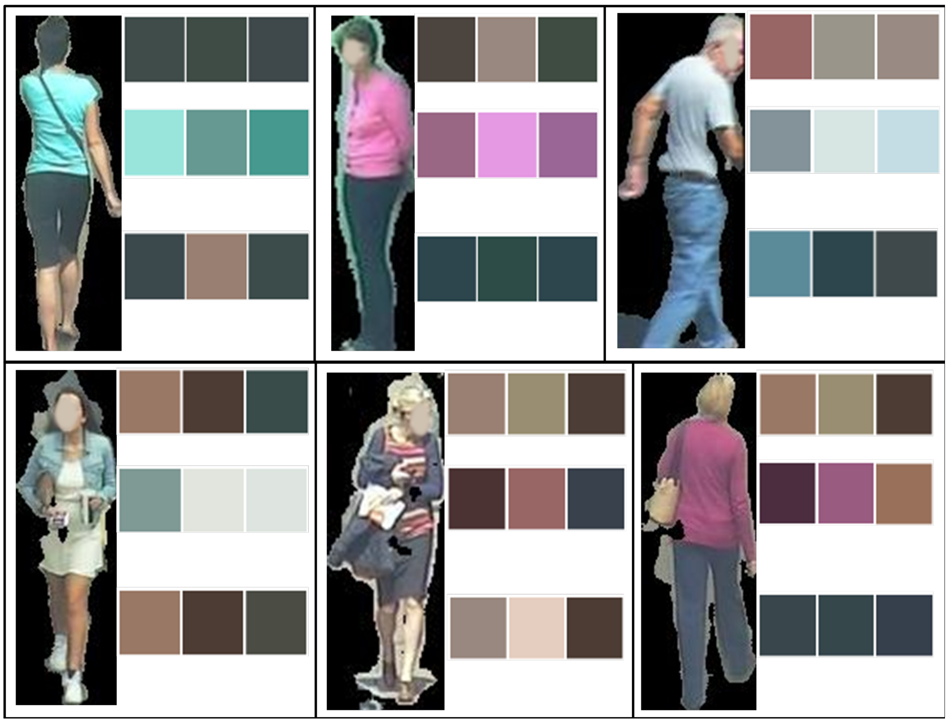


Figure 5.7: Colour recognition results for various human figures. Three dominant colours for each of the three salient body regions have been extracted.

Figure 5.9 illustrates the performance of colour based object retrieval. The colour boxes shown in the 1st column were given as the query colours to retrieve the humans that have their dominant colours similar to the input query colour. The top 4 retrieved results are shown in the figure. It is evident that the proposed framework of colour correction and recognition in conjunction with the use of temporal histogram accumulation has achieved good accuracy not only in terms of extracting the dominant colours but also the retrieval of objects using that colour.

To quantify the performance of the proposed framework, a set of the 247 images were extracted from video frames and divided among 9 basic colour categories namely: Black, Blue, Green, Gray, Red, Pink, White, Yellow and Brown. This is done using subjective analysis and further opinion from research group members was sought to confirm the judgement. All the images of the dataset were used as query images and the values for the True Positives and False Positives were recorded for each class. Finally, the average Recall and average Precision scores were computed to assess the performance of the framework. Table 5.1 shows the average Recall and average Precision achieved by the proposed colour based retrieval framework.

Table 5.2 presents a confusion matrix representing the confusions occurred between various colours. The accuracy of recognition for all the colours is noticeably good. Surveillance videos have poor brightness and contrast and the proposed

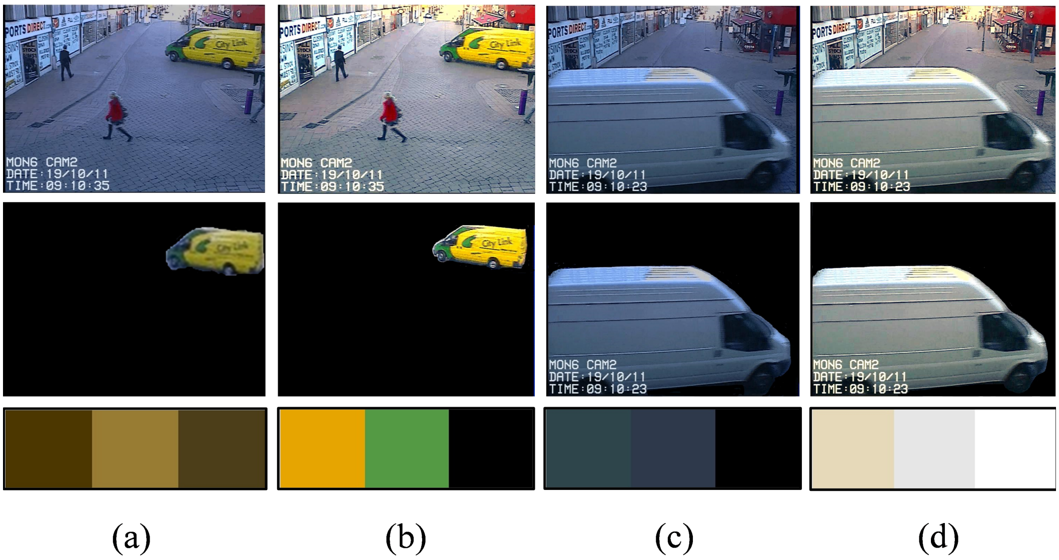


Figure 5.8: Extraction of three dominant colours in vehicles. (a) & (c) are the original video frames along with the estimated dominant colours of the specified objects. (b) & (d) are the colour corrected versions of the frames through the proposed frame work. It can be seen that the extracted colours after correction are accurate.

Table 5.1: Performance parameter values achieved by the proposed colour based retrieval framework.

	Average (%)	
	Recall	Precision
Proposed Method	89.87	93.29

framework has achieved exceptional performance for the extraction and retrieval of objects on the basis of dominant colours. This justifies the positive contribution of the proposed colour correction and enhancement pipeline.

Another set of experiments is performed on two benchmark datasets to test the performance of the proposed colour extraction framework in terms of learning the colour names. The first dataset proposed in [156] comprises of two sets of images collected from two different sources (Google and Ebay). The Ebay dataset is only used here for testing the proposed framework. The Ebay dataset comprises of 523 images of 4 different objects (cars, dresses, pot glasses and shoes). The images in each object category are categorized into 11 colour classes i.e. black, blue, brown, grey, green, orange, pink, purple, red, white, and yellow. The objects present in the images contain the colour of the class to which the image belongs. The pixels in the image are manually labelled and a binary mask of the objects is available with the dataset in order to ease the process of segmenting that object.

The second image dataset named as Google-512 dataset [128] comprises of 5632 images of various objects along with their binary masks. The images are distributed in 11 colour classes similar to the the case of the dataset of [156].

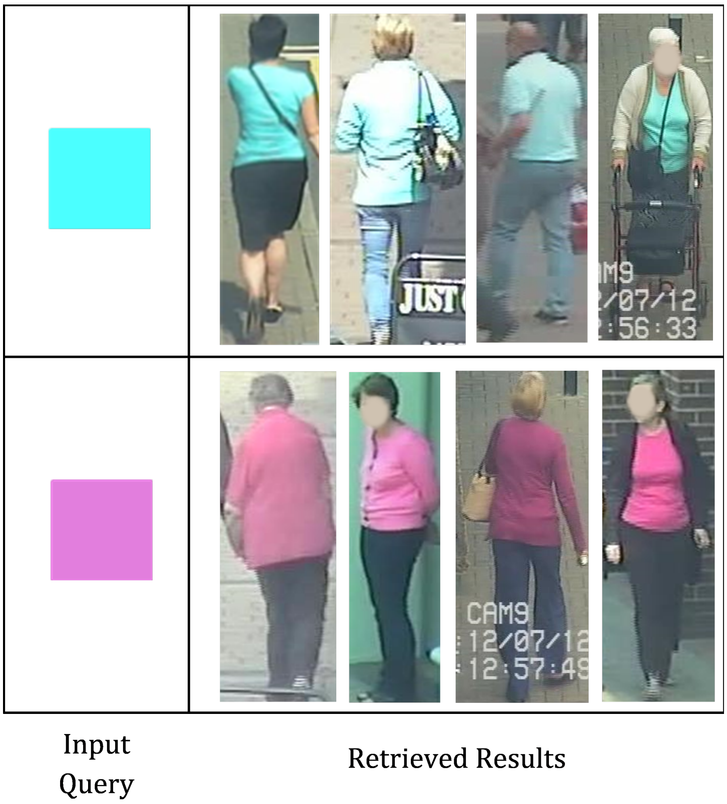


Figure 5.9: The results of the colour based retrieval of human figures. The top 4 results for the input query: Retrieve human figures with specified query colours in torso.

Following [128], this work used the Google-512 dataset for training and Ebay dataset as the test set. The DCDs of the training set are computed using the proposed framework.

While testing, for every test image its DCD is computed and compared with all the DCDs of the training data using the Euclidean distance based similarity measure (see equation 5.13). Finally, the average of the similarity values per class is computed and the test DCD is classified as the colour with which it has the highest average similarity.

The classification results on Ebay dataset are given in table 5.3. Notice that the proposed framework supersede the state-of-the-art results on Cars and Humans categories. This is due the use of weighted strategy where the weights of various colours are altered by using the information about the object. For example in cars, the weight of the black colour is reduced by changing the parameter P in the DCD due to the fact that cars contain various objects i.e. tyres, windscreen etc. that significantly reduce the quantity of pixels present in the car's body and affect the actual dominant colour of the car.

Table 5.2: Confusion matrix for different colour classifications.

	Black	Blue	Green	Grey	Red	Pink	White	Yellow	Brown
Black	0.969	0.056	0.043	0.106	0.023	0.013	0.012	0.034	0.072
Blue	0.005	0.850	0.022	0.032	0	0	0.008	0	0
Green	0.010	0.047	0.873	0.017	0	0	0.005	0	0
Grey	0.016	0.046	0.032	0.81	0.031	0	0.015	0.002	0.008
Red	0	0	0	0	0.893	0.027	0	0	0.001
Pink	0	0	0	0	0.324	0.918	0	0	0
White	0	0	0.022	0.085	0.208	0.042	0.952	0.061	0.014
Yellow	0	0	0.079	0	0	0	0.008	0.896	0.036
Brown	0	0	0	0	0	0	0	0.0618	0.868

Table 5.3: Performance comparison of the proposed framework with other techniques on Ebay dataset.

Method	Cars	Pots	Shoes	Dresses	Total
Proposed	78.93	82.44	89.89	91.35	85.65
SLDA[129]	73.63	80.90	91.82	90.00	84.09
X2 rank [128]	73.63	79.01	92.73	88.18	83.41
PLSA-bg [156]	71.82	83.64	92.73	86.36	83.64
Human Perception [128]	92.73	87.82	90.18	91.99	90.64

5.4.1 Computational Time

The proposed framework is implemented in MATLAB on a standard PC. The average computation time for the proposed colour correction and extraction framework is given in table 5.4. Notice that the noise reduction stage consumes a substantial amount of time. In order to make the proposed framework computationally efficient, the BF based noise reduction scheme may be replaced with conventional Median Filtering (MF) [60] technique. The MF also possesses the edge preserving capability but is not as accurate as BF. The selection of kernel size is also a major issue for various types of CCTV videos. Table 5.5 compares the processing time of various filtering techniques for noise reduction within in the image.

Table 5.4: Average processing time of various stages of the proposed framework.

Stage	Average Time (sec.)
Colour Constancy (Grey World)	0.15
Noise Reduction (Bilateral Filtering)	15.03
Proposed Post Processing	1.03
Total	16.25

Table 5.6 compares the computation time of various colour constancy techniques. Notice that the GW has the minimum computation time among all the

Table 5.5: Average processing time of various filtering techniques for noise reduction.

	Filtering Technique		
	Averaging Filter	Median Filter	Bilateral Filter
Average Time (sec.)	0.07	1.15	15.03

computational colour constancy schemes.

Table 5.6: Average computation time of various Colour Constancy algorithms on a video frame.

Algorithm	Average Time (sec.)
Grey World	0.15
Max RGB	0.17
Shades of Grey	0.225
General Grey World	0.392
Grey Edge	0.415
2nd Order Grey Edge	0.532

5.5 Summary and Conclusion

This chapter has addressed various challenges involved in the accurate colour extraction of objects in surveillance videos for the purpose of annotation and retrieval. The main contribution of the work is an efficient colour correction pipeline that improves and enhances the true colours of the objects by estimating the illumination present in the videos frames. It automatically recognizes and corrects the video frames that are over or under exposed. Apart from colour correction, a weighted strategy is also devised that uses the prior information about the shape of the object and assign weights to the colours that are more or less likely to occur in a particular object of interest i.e. humans and vehicles. The proposed framework has efficiently used the temporal information to improve the confidence of correct colour extraction by computing the colour histograms of objects over various frames.

The proposed framework was tested within a colour recognition scenario as well in a colour based object retrieval scenario. The experiments have been performed not only on publicly available datasets but also on datasets collected by the author. The proposed colour recognition technique has superseded the performance of the state-of-the-art-techniques on benchmark datasets. This chapter also highlighted the issue that the conventional colour constancy schemes alone are not sufficient for accurate colour correction of low quality videos. It was shown that an adaptive post-processing framework is required to enhance the true colours of the objects to aid the process of accurate colour extraction and retrieval of objects on the basis of dominant colours.

Chapter 6

Vehicle Make & Model Recognition and Type Identification

6.1 Introduction

The problem of Vehicle Make and Model Recognition (VMMR) has gained significant importance in recent years due to its potential applications in the field of security and surveillance. In an access control scenario, the VMMR system can add another level of verification apart from existing license plate recognition systems that can be easily deceived with a fake license plate. VMMR systems can be deployed on toll plazas to automatically issue relevant tickets by recognizing vehicle types. The main emerging application in surveillance and law enforcement is to search a specific vehicle in the CCTV feeds. The license plate recognition, nonetheless, requires specialized high resolution/high speed cameras and in general cannot operate on conventional CCTV video feeds. In such applications the number plate based identification is, therefore, not reliable. The recognition of vehicle's make and model using the appearance information can, not only add another level of security but also is able to operate on images captured from conventional CCTV cameras.

The work presented in this chapter is significantly based on the proposed framework in [45]. The proposed framework is inspired by the observation that the problem of VMMR using frontal view is similar to human facial recognition problem. In the detailed review of existing VMMR techniques in chapter 2 (section 2.5), it is noticed that none of the existing techniques have exploited this observation to solve the said problem. This chapter therefore exploits the state-of-the-art facial recognition technique [176] and object recognition methods [20, 142] to propose a novel framework that reliably recognizes the make and model of a vehicle from images/videos in the presence of wide imaging variations. The strong inter-class similarity between various makes of vehicles and significant intra-class variability among the models of the same make poses a challenge for accurate recognition of vehicles.

6.2 The Starting Point of the Research

The work in this chapter is inspired by the work of Liao et al. [136] who proposed the use of patch level representation of facial images to generate an over complete dictionary of visual words.

Recently proposed approaches in object recognition [20, 142] have focused on developing the image representations by dense computation of existing off the shelf feature representation techniques. The dimensionality of densely sampled features is reduced by employing mid-level feature encoding techniques. These techniques aim at reducing the dimensionality of dense feature representations while establishing a meaningful discrimination among them. The mid-level feature representation has been exploited in this work to generate an over complete dictionary of visual words from the training images of vehicles. All the vehicle make and model classes are represented using a distinct combination of these visual words.

While representing images as occurrences of similar visual words from a large dictionary is typically found in Bag of Visual Words (BoVW) based approaches [29], carrying it out on the patch level was recently proposed by Liao et al. [136] in the context of face recognition. In [136], encoding each patch in terms of a sparse representation of an over complete dictionary is achieved by employing the Sparse Representation based Classification (SRC) [176]. Since a large dictionary of patches computed from all the training images typically comprises of millions of patches, finding the sparse coefficients for each test patch in terms of this dictionary requires solving a highly computational extensive optimization problem for each test patch. As opposed to this, the main advantage and contribution of the proposed VMMR approach is to achieve the same level of representation (in terms of visual words) by employing the learned Mid Level Representation (MLR) and an efficient encoding scheme. This work thereby draws a useful and intuitive dual between such sparse representation based schemes and the MLR based encoding of densely computed features.

6.3 The Proposed System

6.3.1 Overview

The block diagram of the proposed VMMR framework is shown in figure 6.1. The proposed framework starts by locating a vehicle in an image or a video frame. A part of vehicle's front is selected as RoI. Next, the visual words are computed from the Region of Interest (RoI) and classified using the comparison with the lexicon of visual words of training images.

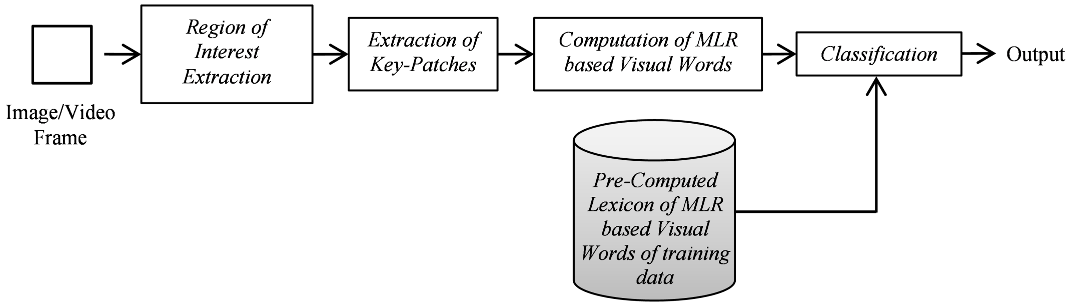

Image/Video Frame
Region of Interest Extraction
Extraction of Key-Patches
Computation of MLR based Visual Words
Classification
Output
Pre-Computed Lexicon of MLR based Visual Words of training data

Figure 6.1: Block diagram showing key stages in the proposed vehicle make and model recognition framework.

The performance of existing vehicle make and model recognition techniques significantly relies on a pre-processing stage where the images are tailored and close-fitting RoI selection is performed to reduce the effect of background noise in extracted feature representation. The strength in the proposed framework is its relatively low reliance on RoI extraction stage. Though a slight pre-processing and rough RoI estimation is essential as an initial step in the proposed framework, unlike most of the existing techniques, it does not require a precisely registered, de-skewed and tightly cropped RoI to produce an accurate recognition result.

The experiments have shown promising results and potential strength of the devised methodology. The main contribution here is to show that the over complete lexicon of MLR computed from dense feature descriptors sampled around stable key-points in an image creates an efficient image representation that reliably copes with background noise and performs better than existing techniques in terms of accuracy for vehicle make and model recognition. The proposed framework offers a fast classification (see section 6.4.6) of vehicle data. Another advantage of proposed method is that it does not require a retraining if new vehicle make and model classes have to be incorporated to the dataset. Therefore, the capability of the system can be easily enhanced to new vehicle classes. The key stages of the proposed framework are separately explained in detail in the following sections.

6.3.2 Reference Detection and RoI Extraction

In this work, the focus is to recognize the vehicle's manufacturer and model using the frontal view of the cars. The first significant step therefore is to accurately localize the vehicle in the scene image and extracting a RoI. Though, the proposed feature representation scheme is accurate even in the presence of unwanted background information, yet the RoI selection is important due to the fact that a specific part (region containing grills, headlights and some portion of bumper) of the vehicle's front contains useful information that discriminates a particular model from others.

The schemes proposed in the literature have mainly used two different methods to detect a vehicle and extract its RoI. The first and most widely used method to detect a vehicle [112, 123] is based on locating its license plate and uses its geometric properties as a reference to extract an area around it as the RoI. The method is extremely reliable and computationally fast. The reason being the license plate contains highly discriminative visual features compared to the background and can therefore be easily detected even in cases where limited angular variation and occlusion is present.

The second vehicle detection scheme[59]uses symmetric appearance features to detect a vehicle's front and then select a RoI from it. The method exploits the fact that half of the vehicle's front is exactly the mirror image of the other half if the vehicle's front is divided into two parts using a vertical line. The technique is reliable in poor lighting where the license plate is hard to detect. However, the main disadvantage of this technique is its failure in case of occlusion and stroppy camera angle. In case of angular view, this method detects the vehicle but fails to centre the region well because of the absence of a reference point. This result in a significant proportion of background information encompassed as part of RoI and affects the recognition accuracy in a negative way. Therefore, in this research the license plate detection based approach has been adopted along with the incorporation of some novel modifications to reliably extracting the RoI. The proposed method is simple and robust to various noise factors that the symmetric appearance based method fails to cope with.

Given an image with a vehicle present in it, a basic license plate detector [124] is applied to locate license plate in the image. The license plate detector exploits various geometric and gradient features to locate the license plate in the image. It applies various morphological operations to enhance horizontal and vertical edges, since the density of horizontal and vertical edges is significantly high in a license plate due to its rectangular boundary and presence of numbers. After enhancement, the connected components are extracted and passed through two classification stages to remove all non-plate regions. The first stage analyses the regions on the bases of geometrical features. These features include height, width, aspect ratio and edge density of each region. The thresholds for these features have been empirically adjusted and the candidate regions that fail to fulfil the established criteria are rejected at this stage. The candidate regions that pass initial screening are then fed into the second stage where they are classified on the basis of their shape features. A significant variety of shape features have been proposed in literature for classification of objects and patterns in a number of scenarios. In this work, the most commonly used HOG features have been utilised to represent candidate regions. The classification process uses precomputed HOG features for a brief set of pre-cropped training images of license plate and non-license plate regions. A Support Vector Machine (SVM) classifier is learned using the training data to classify the candidate regions in the test image.

Once the license plate is detected, the image within the estimated rectangular bounding box is used for applying a post processing procedure that straightens the license plate image and removes any skew. The detected bounding box is horizontally and vertically aligned with the vehicle image. However the license plate within it is not always horizontally straight and may have some degree of skew or angular variation. This variation is removed using the proposed two-step post processing procedure:

1. The first step reduces the in-plane rotation within the image. The edges of the license plate image within the bounding box region are computed using a Canny edge operator. Next, the line segments within the edge map are extracted using the Hough transform. After extracting all the line segments, each line segment is analysed based on two criteria: 1. Length (L), 2. Angle (θ). The line which is the longest and has the smallest angular difference to the horizontal axis (0 degree) is selected as a part of the boundary of the license plate. This line segment is used as a reference to remove the angular variation within the image. Consider a line segment l estimated as the boundary part of the license plate having length L_l and angle θ_l. The angle θ_l is used to remove the angular variation within the image. The vehicle image is rotated at an angle of $-\theta_l$ degree. The process effectively removes any in-plane rotation of the vehicle within the image. Figure 6.2 illustrates the process on a vehicle image with 5 degrees of in-plane rotation.

2. The second step is applied to the image that is acquired after reducing the in-plane rotation in the first step. This step aims at reducing the skew within the image that usually occurs if the image is captured from a slightly varying angle than the perfect frontal view of the vehicle. The four corners of the

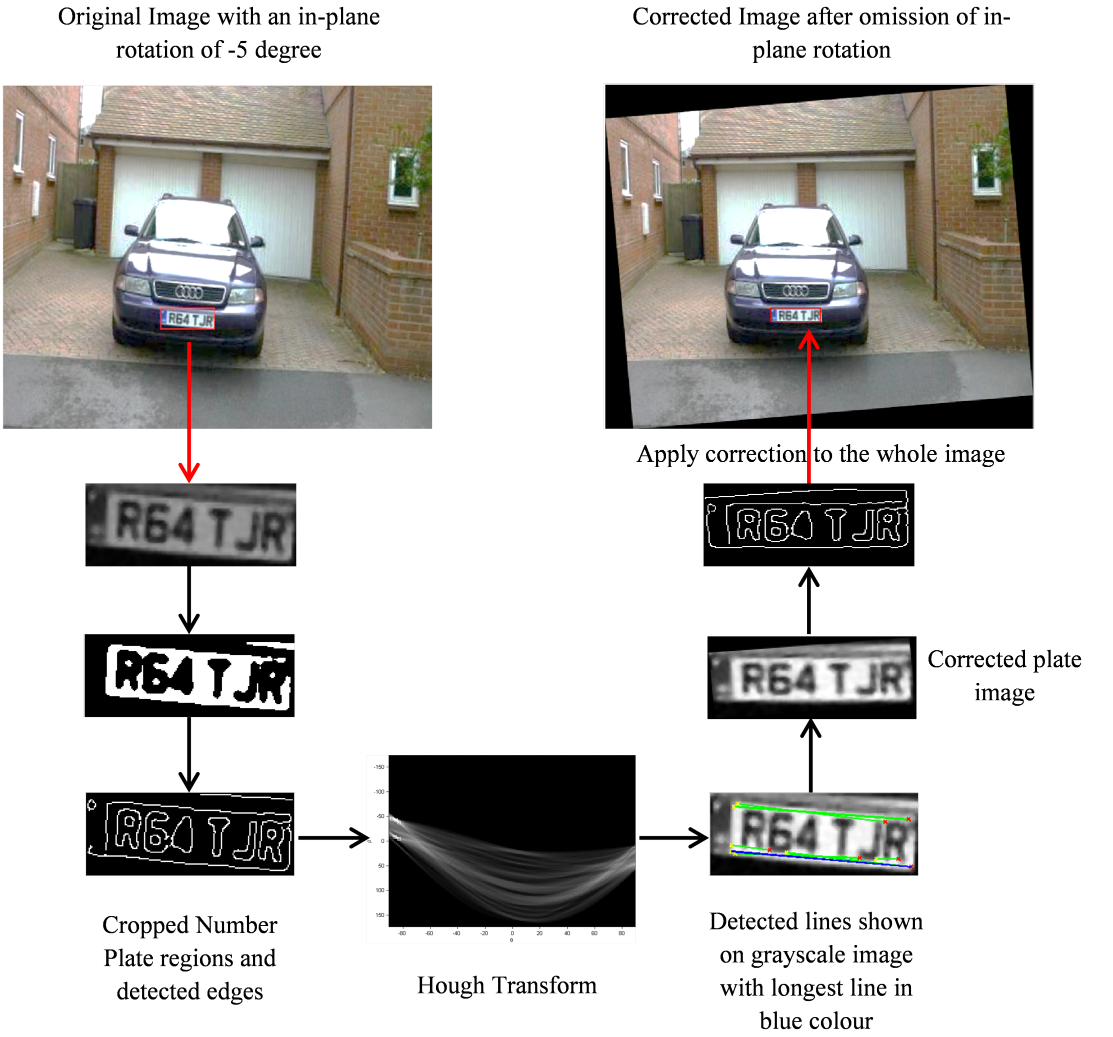

Original Image with an in-plane rotation of -5 degree
Corrected Image after omission of in-plane rotation
Apply correction to the whole image
Corrected plate image
Cropped Number Plate regions and detected edges
Hough Transform
Detected lines shown on grayscale image with longest line in blue colour

Figure 6.2: Correction of in-plane rotation using the horizontal edges in the license plate image.

license plate are used as reference points to register (via affine transform) the under query license plate to a reference license plate image, which is perfectly horizontal and has no effect of skew.This procedure removes any skew, scale difference and in-scene rotation that can potentially affect the precise RoI extraction. The transformation matrix obtained from the reiteration of license plate is applied to the whole vehicle image prior to RoI extraction. Figure 6.3 shows an image before and after image de-skewing and angular variation correction.

The aim in RoI selection is to reduce the background information that acts as noise for vehicle make & model recognition. Consider c as the centre and W as the width of the detected license plate, the RoI is extracted using the criteria given in equations 6.1 and 6.2.

$$x_1 = c - 1.4W, y_1 = c - 0.8W \tag{6.1}$$

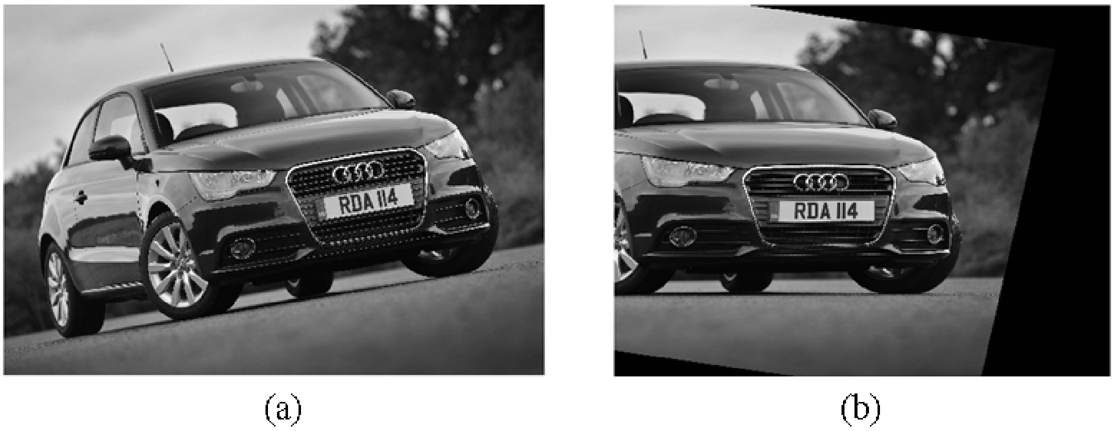


Figure 6.3: Correction of image using affine transform based image registration. (a) Original Image with in-plane rotation and skew. (b) Affine warped image containing vehicle aligned with horizontal axis.

$$Width_{RoI} = 2.8W, Height_{RoI} = 1.2W \tag{6.2}$$

Here, x_1 and y_1 are the coordinates for the top left corner of the RoI bounding box. Figure 6.4 illustrates the RoI extraction criteria. The criteria are empirically selected with the aim that vehicle information is not missed in an effort to attain a tight bounding box. Therefore, the aim is not to miss any vehicle information even if some background is also selected as part of the RoI. Since, one of the applications of a VMMR is to recognize the type of the vehicle, the RoI selection needs to be global enough to work correctly on bigger vehicles like buses and trucks etc. as well as small vehicles like cars. Therefore, the RoI selection criteria represented by eq.6.1 and eq. 6.2 never clips the vehicle information; however, it always selects some part of background as RoI which is easily accomodated by the strong patch based feature representation scheme.

The extracted RoI is resized to 128-by-256 pixels. The size is selected empirically and the aspect ratio is maintained at 1:2 which usually is the case in any vehicle's RoI in this work.

6.3.3 Extraction of Key-Patches

The RoI of vehicle image is represented by computing the feature descriptors on the patches that are extracted around stable key-points within the image. The process starts by the detection of stable key-points within the image. The key-point detection is performed using a Difference of Gaussians (DoG) [88] based key-point detector applied at 3 different scales of the image. The key-point detector as presented in [88] first searches for potential interest points over all scales and image locations to identify such interest points that are invariant to scale and orientation.

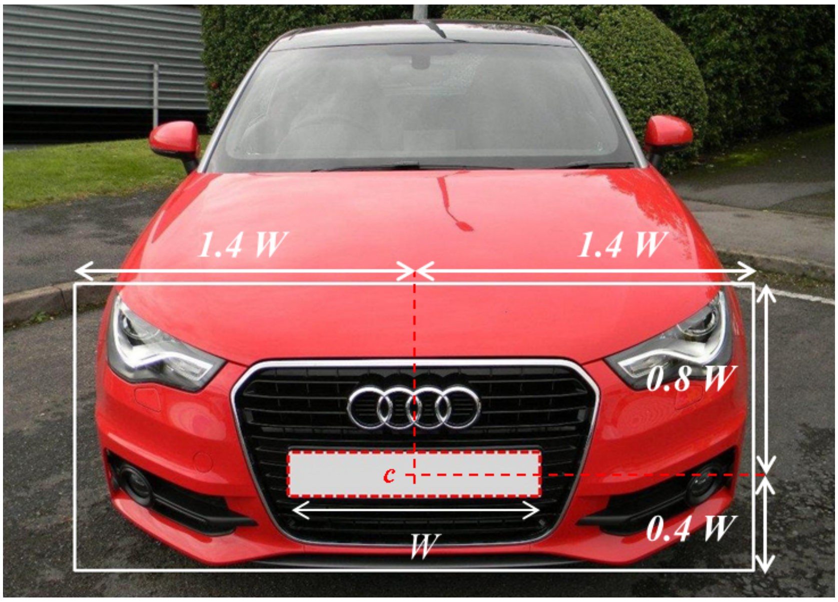


Figure 6.4: Selection of RoI in a vehicle image. Here c is the centre of the number plate having width W.

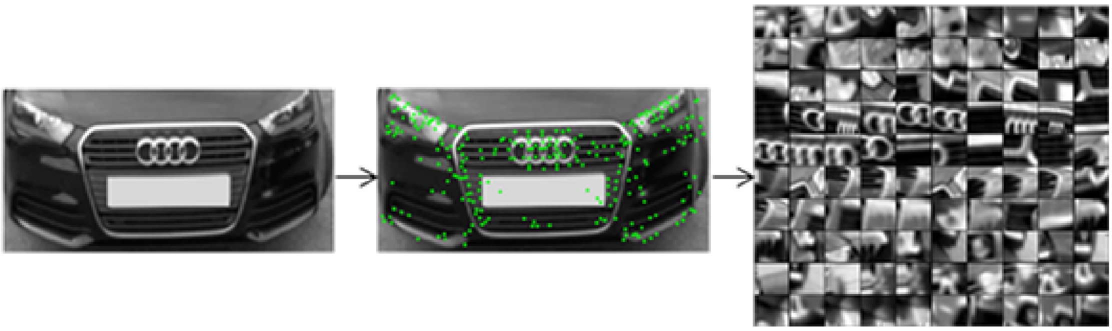


Figure 6.5: Detected key-points and the extracted key-patches around stable key-points. For illustration purpose, first 100 key-patches are shown here.

It then fits a model to determine the location and scale at each candidate location and finally selects the key-points based on the measure of their stability. After detecting stable key-points within the image, a 41-by-41 pixels patch is extracted around each key-point as a stable key-patch. In case of 128-by-256 pixel images, the extracted number of patches varies between 150 and 300 for different training examples. Figure 6.5 shows an example where the key-patches around stable key-points are extracted from a vehicle's image. For the purpose of illustration, only first 100 key patches are shown here.

6.3.4 Computation of Visual Word

The key-patches extracted in the previous stage are then used to compute the MLR based visual words. One visual word per patch is computed. Each image is represented by the collection of visual words computed from its key-patches.

The visual word computation procedure begins with the computation of dense

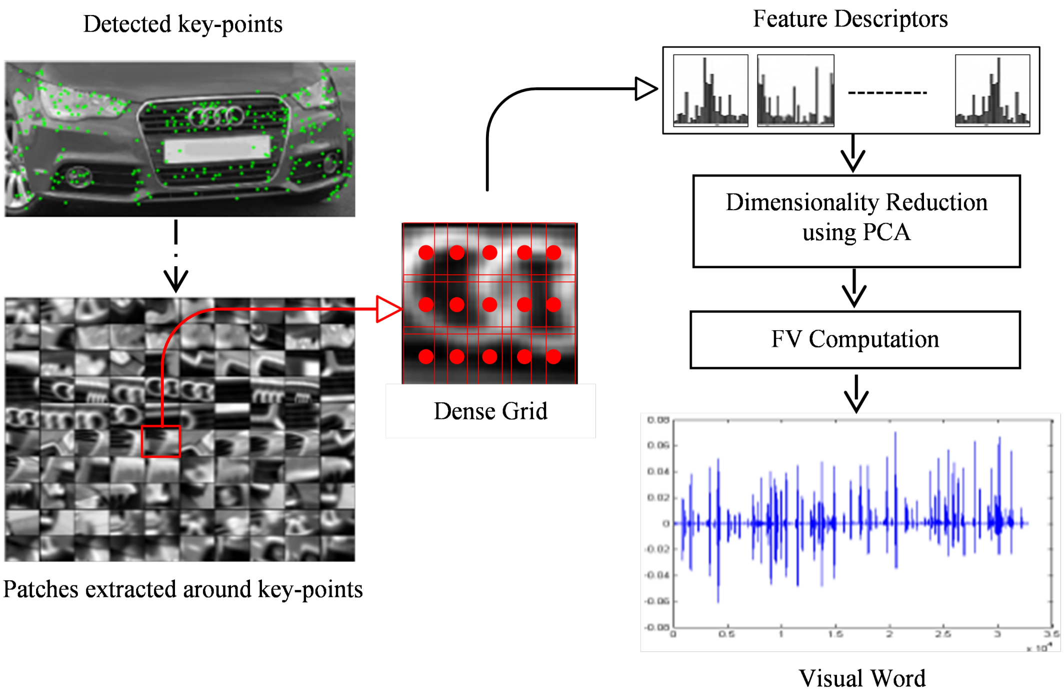


Figure 6.6: Visual word computation from dense feature descriptors.

SIFT features at sampled locations within each key-patch at three different scales. After the feature descriptors are computed, MLR is computed to aggregate the densely computed feature descriptors into a single discriminative description referred as visual word. In this work, the non-linear Fisher Vectors (FV) encoding based on the diagonal covariance matrix of the Gaussian Mixture Model (GMM) is used as MLR scheme. The mathematical procedure for the computation of Fisher Vector encoding is explained in section (3.12). Figure 6.6 illustrates the process of visual word computation from a key-patch.

Let, for a training image I the extracted key-patches be $p_1, p_2, ..., p_n$. For every patch p_q, where $q = 1, 2, ..., n$, feature descriptors $f_1, f_2, ..., f_M$ are computed using a dense overlapping grid, where M is the number of windows in the dense grid. The resulting output contains M no. of feature vectors per patch, each of length d.

In this work, 24-by-24 pixel windows with a stride of 6 pixels at three different scales of the image have been used to extract dense SIFT features. The SIFT feature vector for each window is computed that results in $M = 41$ and $d = 128$. In the step that follows, FV encoding is performed to aggregate the densely computed feature descriptors of the patch p_q into a single vector representation referred as a visual word and denoted by V_q.

The dimensionality of each FV is $2Gd$ where G is the number of Gaussian centres (in this case, $G = 256$) in the GMM and d is the dimensionality of each feature vector in the dense feature description of each patch. The dimensionality of FV is significantly high and directly depends upon the dimensionality of feature vectors. If the dimensionality of feature vectors is reduced, the dimensionality of FV is automatically reduced. In order to ensure this, the dimension of the densely computed SIFT feature vectors from each patch is reduced from 128 to 64 through Principle Component Analysis (PCA) [153]. The dimensionality reduction of feature vectors results in a significant reduction in the dimensionality of FV from 65536 to 32768 (for $G = 256$). The spatial information of patches is not incorporated in the feature description. This is due to the fact that the key-patches work as a vocabulary and the location of key-patches in different images is different; therefore, experiments revealed that the incorporation of spatial information cause degradation in performance.

6.3.5 Construction of Lexicon from Visual Words

A visual word is computed from every patch of the training image. The pool of visual words from the vehicle images belonging to same class constitutes a sub-lexicon. The visual words in a sub-lexicon are kept under the label of that vehicle class to which those visual words represent. The collection of sub-lexicons from all training classes form one comprehensive lexicon containing the over complete representation of each vehicle class with the help of visual words. Figure 6.7 illustrates the process for the construction of lexicon using the visual words of each vehicle class that comprises sub-lexicons from all training class.

For n number of key-patches within an image I, n visual words are achieved. The visual words from all the images of the same training class t, where $t = 1, 2, ..., T$, form a sub-lexicon S_t, of size $P_L \times L$, where L is the total number of images in t and P_L is the total number of key-patches found in L images.

All sub-lexicons from a total of T classes of the training set collectively form a complete lexicon X, generating an over complete description space of the T classes (eq. 6.3) because of significantly large number of visual words representing each class.

$$X = [s_1, s_2, ..., s_T] \tag{6.3}$$

6.3.6 Classification

The classification of the test image starts by vehicle detection and RoI extraction using the method explained in section 6.3.2. The selected RoI is resized to 128-by-256 pixels and fed into the stage where visual words from that image are computed. In the next step, each visual word from the probe image is matched against the visual words within each sub-lexicon using Cosine Similarity (CS) metric.

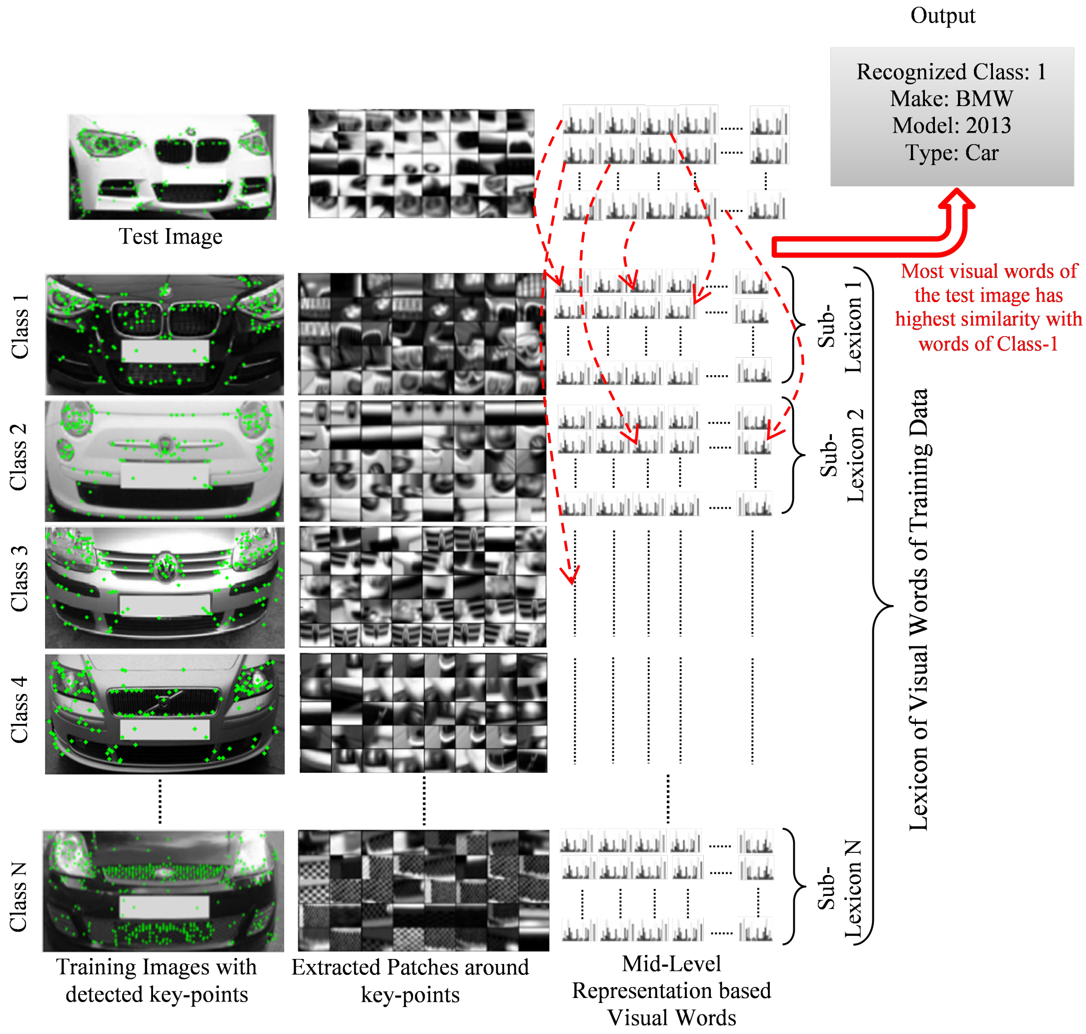

Output
Recognized Class: 1
Make: BMW
Model: 2013
Type: Car
Test Image
Most visual words of the test image has highest similarity with words of Class-1
Class 1
Class 2
Class 3
Class 4
Class N
Sub-Lexicon 1
Sub-Lexicon 2
Sub-Lexicon N
Lexicon of Visual Words of Training Data
Training Images with detected key-points
Extracted Patches around key-points
Mid-Level Representation based Visual Words

Figure 6.7: Computation of Lexicon of visual words containing visual words from each training class.

Let $j-th$ sub-lexicon contains v number of visual words; then for each visual word of the test image, v scores are attained. If there are n patches in the test image, then a v-by-n dimensional matrix is achieved. The maximum similarity score for each test patch is selected using max-pooling. This results in a single similarity score per patch. The sum of all these scores is recorded as the similarity score of the test image with the $j-th$ class. In this way, the similarity score with all the classes is computed and finally the test image is categorized to that class for which it has the highest similarity score.

6.4 Experiments and Results

6.4.1 Image Datasets

The proposed framework is tested on two publicly available image datasets [186, 27]. The first dataset [186] comprises of 300 front view vehicle images divided among 25 classes where each vehicle class contains at least 11 images. As in [186], 100 training images (4 per class) have been used for lexicon construction and all the remaining images have been used as test images. The second dataset named as PANMMVR dataset [27] comprises 166 high quality frontal vehicle images for training. The images have been captured in different scenarios: streets, outdoor and indoor car parks. The test database contains 708 outdoor frontal vehicle images in different lighting conditions and with lower resolution. The authors mentioned that the images were recorded using two video camcorders with an optical zoom.

Loughborough Cars Dataset: Another contribution of the research presented in this thesis is an updated comprehensive dataset of vehicle images names as Loughborough Cars (LC) Dataset. The dataset has been collected to evaluate the proposed framework and to provide a benchmark for future researchers. Since, the original LC dataset was compiled in the year 2008 and a large number of new vehicle models have been released since, the need for an up-to-date dataset was felt. The images for the new datasets have been collected mainly from public car sales websites. The updated LC dataset is now a technically challenging collection of images since the resolution, size and quality of images greatly vary as they have been captured by various users of car sales websites with their personal mobile phones and camera devices. Apart from variability in quality, the images have high diversification in terms of camera angle as can be seen in figure 6.8. A total of 1537 (front view) car images have been collected and divided in 75 make and model classes. Six images from every class have been used for training. Conclusively, 450 training images and 1087 test images have been used for the experiments. The total number of car images for each make and model are given in table 6.1.

6.4.2 Video Datasets

The capability of the proposed framework is also tested on the videos. For this purpose, two comprehensive video datasets have been collected. The videos for the first dataset were recorded using a high definition camera with a perfect frontal view of the vehicles. The video clips were captured at different times of the day

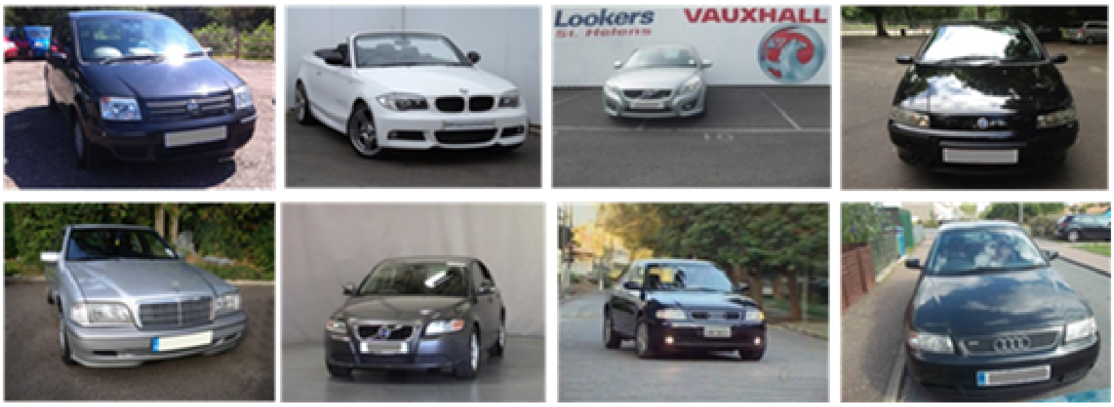


Figure 6.8: A few images from challenging LC dataset.

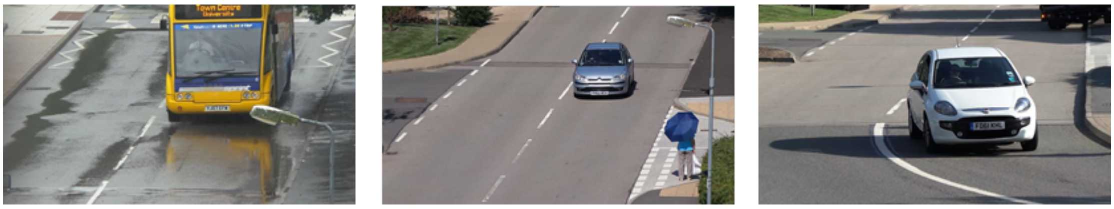


Figure 6.9: A few video frames from the high definition video Dataset. Notice that the videos have been recorded in different weather conditions and from different distances.

on a less busy road. The dataset contains 7613 video frames of 1280 × 760 pixels and distributed into 97 classes. Figure 6.9 shows a few video frames from the high definition video dataset.

The second and most challenging dataset is compiled using real surveillance videos acquired with the help of Loughborough town's CCTV surveillance centre. The video clips acquired have been recorded by a CCTV camera installed on a roadside within Loughborough. The video frames of different vehicles have been manually separated for testing. The dataset comprises of 3746 video frames of 720 × 576 pixels each divided amongst 97 classes. Since, the video clips have been recorded at different times, the quality, viewing angle and lighting significantly varies in different videos. Figure 6.10 shows a few video frames from the CCTV dataset.

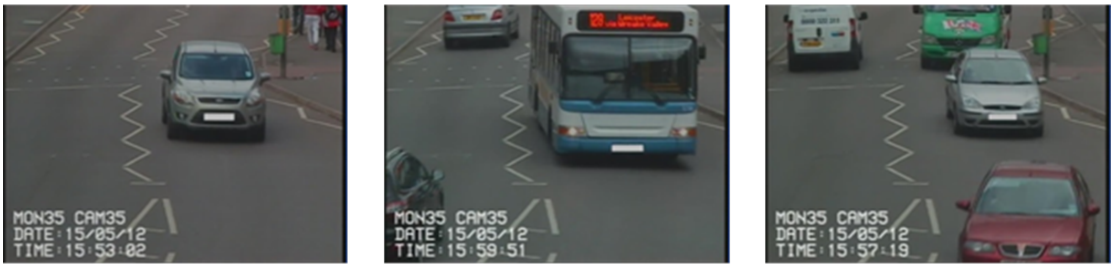


Figure 6.10: A few video frames from the challenging real CCTV vehicles Dataset. Notice that the video frames have significantly low quality where minute vehicle features are not sufficiently distinctive.

Table 6.1: Number of images in each make and model class of LC dataset. The Model column displays the model name and the year in which that particular model was first released.

Make	Model	No. of Images
Audi	A1, 2013	18
	A1, 2013	16
	A3, 1996	8
	A3, 2003	10
	A3, 2005	17
	A4, 1994	10
	A4, 1997	10
	A4, 2001	31
	A4, 2004	30
	A4, 2008	27
	TT, 1998	21
	TT, 2005	28
BMW	11, 2004	27
	11, 2011	10
	336, 1990	14
	346, 1998	18
	390, 2005	22
	330, 2012	19
	539, 1995	24
	560, 2003	27
	510, 2009	30
Citroen	C3, 2002	28
	C3, 2009	27
Ford	Cmax, 2003	8
	Cmax, 2007	24
	Cmax, 2010	20
	Fiesta, 1995	21
	Fiesta, 1999	9
	Fiesta, 2003	20
	Fiesta, 2008	28
	Focus, 1998	19
	Focus, 2005	19
	Focus, 2010	17
	KA, 1996	20
	KA, 2008	24
	Mondeo, 2000	20
	Mondeo, 2007	13
Fiat	500, 2007	29
	Panda, 2003	13
	Punto, 1993	14
	Punto, 1999	37
	Punto, 2005	24
Honda	Jazz, 2001	26
	Jazz, 2008	23
	Civic, 1992	8
	Civic, 1996	33
	Civic, 2001	21
	Civic, 2006	33
	CRV, 2002	32
	CRV, 2007	29
Mercedes	C, 1993	24
	C, 2001	27
	C, 2007	22
	E, 1996	25
	E, 2003	20
	E, 2010	11
Nissan	Juke, 2010	16
	Micra, 1993	18
	Micra, 2003	32
	Micra, 2010	17
	Qashqai, 2007	17
	Qashqai, 2010	15
Volkswagen	Golf, 1993	16
	Golf, 1997	27
	Golf, 2003	26
	Golf, 2008	17
	Passat, 1996	16
	Passat, 2001	18
	Passat, 1996	15
	Passat, 2001	15
Volvo	C30, 2006	14
	C30, 2010	20
	S40, 1995	15
	S40, 2004	21
	V50, 2000	18

Table 6.2: Performance comparison of various approaches on three different image datasets.

Dataset	Recognition Accuracy (%)			
	Zafar et al. [186]	Sarfraz et al. [127]	Clady et al. [27]	Proposed Method
Zafar et al. [186]	84.00	94.00	-	97.60
PANMMVR [27]	-	-	94.00	95.15
LC Dataset	-	78.83	-	84.31

Note that here the training image set (4 images per class) is formed for 97 classes present in test video sequences both from LC dataset and the images obtained from the web for the classes that are not present in LC dataset. The license plate detector mentioned in section 6.3.2 is applied on each frame and the performance was estimated on the basis of number of frames classified correctly in each class.

6.4.3 Performance Metric

The performance metric used to evaluate the proposed framework is the recognition accuracy (%). The recognition accuracy (%) measure is defined as:

$$Rec.Acc.(\%) = \frac{TotalCorrectClassifications}{TotalTestImages} \times 100 \tag{6.4}$$

6.4.4 Results and Comparison

Table 6.2 presents the recognition accuracy of the proposed frame work. The proposed framework is compared with the techniques presented in [186] and [127]. It is evident from the recognition scores that the proposed approach performs better than the published state-of-the-art on this dataset. The methodology presented in [127] is also tested on the LC dataset. It is observed that the recognition accuracy of [127] on LC dataset is significantly lower than on the dataset in [186] owing to the challenging nature of the images and large number of classes (makes and models). The proposed framework on both of these datasets has proved to be very efficient in capturing fine inter and intra class variations among different makes and models of the vehicles.

Figure 6.11 presents the class-wise recognition score and confusion matrices for the LC, Zafar [186] and PANMMVR [27] datasets that were used for the experiments. Figure 6.11a presents the rank curve on LC dataset, it can be seen that the recognition accuracy improves from 84% to 95% for first three matches (rank-3). This indicates the potential of operating the proposed framework in

Table 6.3: Recognition accuracy of proposed framework on video datasets.

Dataset	Recognition Rates (%)
HD Video Dataset	88.09
CCTV Video Dataset	78.48

practical scenarios like on-road vehicle identification and access control. For the Zafar [186] dataset, it can be seen in figure 6.11d that the confusion has occurred in only two classes; images in all other classes have been recognized with 100% accuracy. In the same way, the confusion matrix for PANMMVR [27] in figure 6.11f shows that most of the classes have been recognized with 100% accuracy. A few successful recognition examples on LC dataset are shown in figure 6.12.

The performance of the proposed framework on video datasets is given in table 6.3. The proposed framework achieved a recognition rate of 78.48% on Low quality surveillance videos and 88.09% recognition rate on high quality videos. Figure 6.13 presents the class-wise recognition score and confusion matrices for both video datasets that were used for the experiments. A few visual results on video frames are shown in figure 6.14 and figure 6.15.

The confusion in recognition occurs in various classes as can be seen in figures 6.11 and figure 6.13. The reason being the LC dataset and video datasets have been compiled while separating the vehicles even for minor shape lifts. This means that the classes are not only separated on the bases of makes and models but also on the bases of years, special editions and minor shape alterations. For instance Ford Fiesta 2001 model is confused with Ford Focus 2002 model; similarly BMW 346 is confused with BMW 539(see figure 6.16). The confusion in these classes is well expected and difficult to avoid because subtle difference exists in these classes as being the same make. It may be better to group models that are highly similar into large super-classes that contain several similar models. However, it is worth noticing that the manufacturers of these cars have been recognized with full accuracy.

6.4.5 Effect of varying PCA and GMM parameters on System's Performance

One key step in Fisher Vector computation is the learning of GMM. In order to compute GMM, patch data from images of 10 training classes have been used. The selection of classes is not important. First 10 training classes have been used in this work. The same learned model is used for the experiments with other datasets. The high recognition results show the strong generalization ability of the learned model and eliminate the need of retraining of model for every other

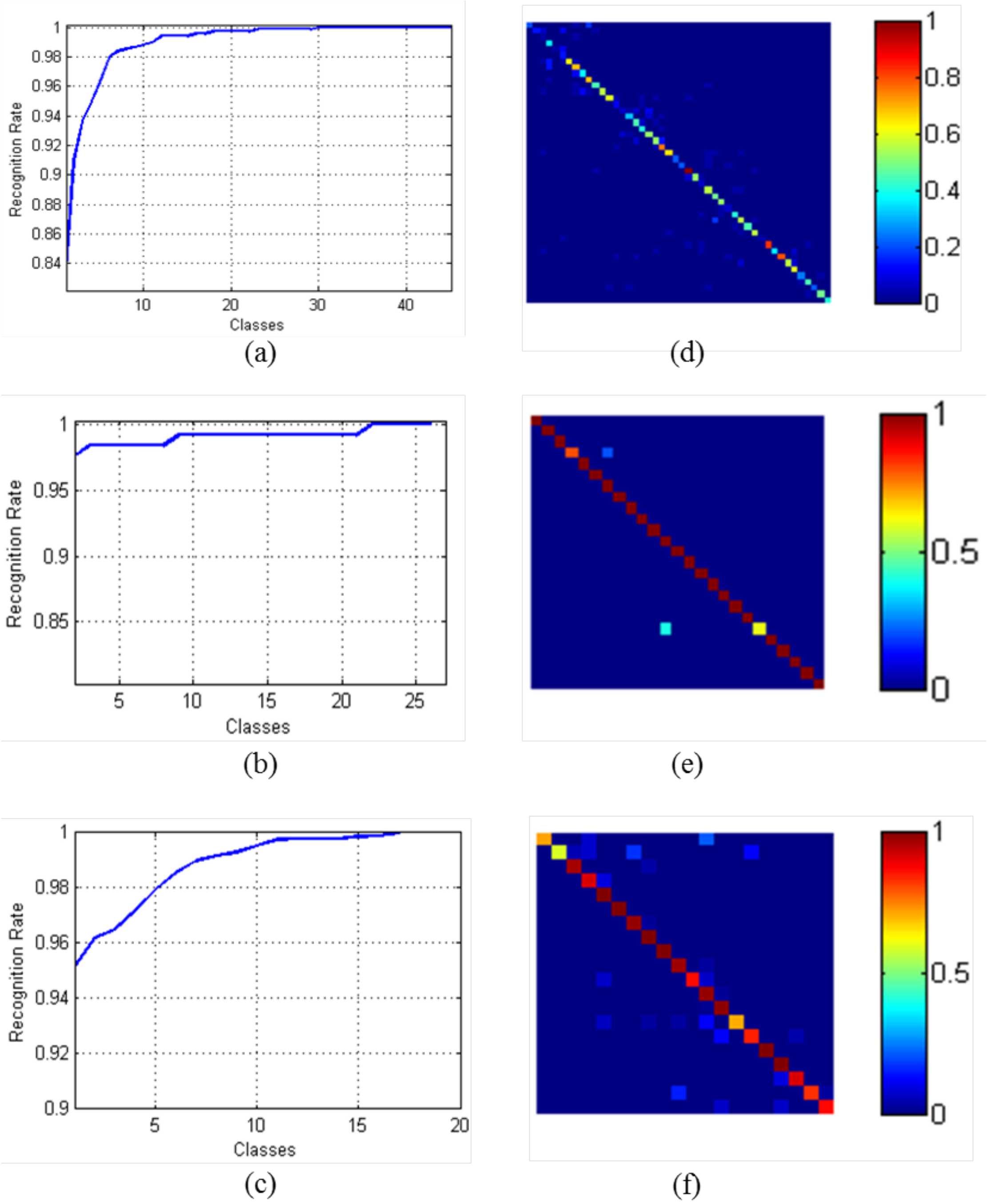


Figure 6.11: Confusion Matrices and Rank-wise recognition accuracy for three datasets. (a) Rank-wise recognition accuracy on LC dataset. (b) Rank-wise recognition accuracy on dataset of [186]. (c) Rank-wise recognition rate on dataset of [27]. (d) Confusion Matrix showing classification accuracy across each class in LC dataset. (e) Confusion matrix showing classification accuracy of proposed framework on the dataset of [186]. (f) Confusion Matrix showing classification accuracy across each class in dataset of [27].

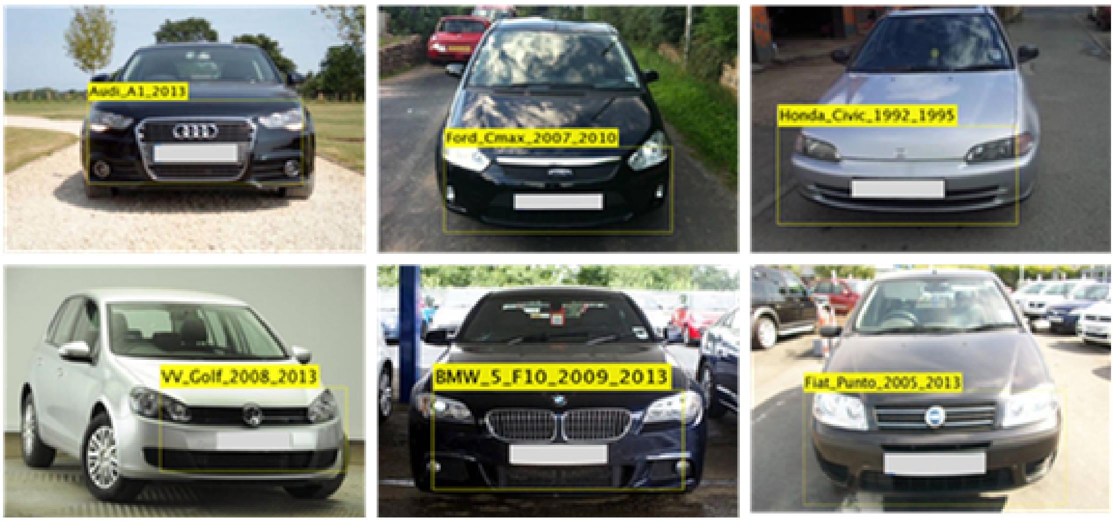


Figure 6.12: Successful recognition results on the images of LC dataset.

dataset. This highlights one of the key advantages of the approach that it does not need retraining for new dataset which in practical systems is otherwise needed. The selection of the number of Gaussian distributions for GMM learning is also important and recognition results vary while changing the number of Gaussian distributions. The experiments were conducted with different number of Gaussian distributions (16, 32, 128, 256, 512, 1024 and 2048) in GMM and with various choices of reduced dimensionality values (16, 32 and 64) of feature vectors using PCA. It was estimated that the GMM learned with 256 Gaussian distributions with dimensionality reduction of feature vectors up to 64 offers a good trade-off between the upper limit on encoded vector dimensionality and recognition ability. Figure 6.17 shows the recognition accuracy of the proposed method using various combinations of feature dimensionality and number of distributions in GMM. Note that the combination of the GMM learned with 256 Gaussian distributions and dimensionality reduction of feature vectors up to 64 achieves the highest recognition accuracy.

6.4.6 Performance of the Proposed System using LESH Features

The experiments were also conducted using the LESH [125] features of patches in the same way as it was done using SIFT. The results revealed the strength of the visual word based representation approach. The experiments performed with 512-dimensional LESH features produced recognition accuracy of 73.83% on LC dataset which is not as high as it is with SIFT features, however, it may still be considered as a good performance. The reason for a decline in the performance is the fact that LESH feature description is not rotation invariant. Therefore, it

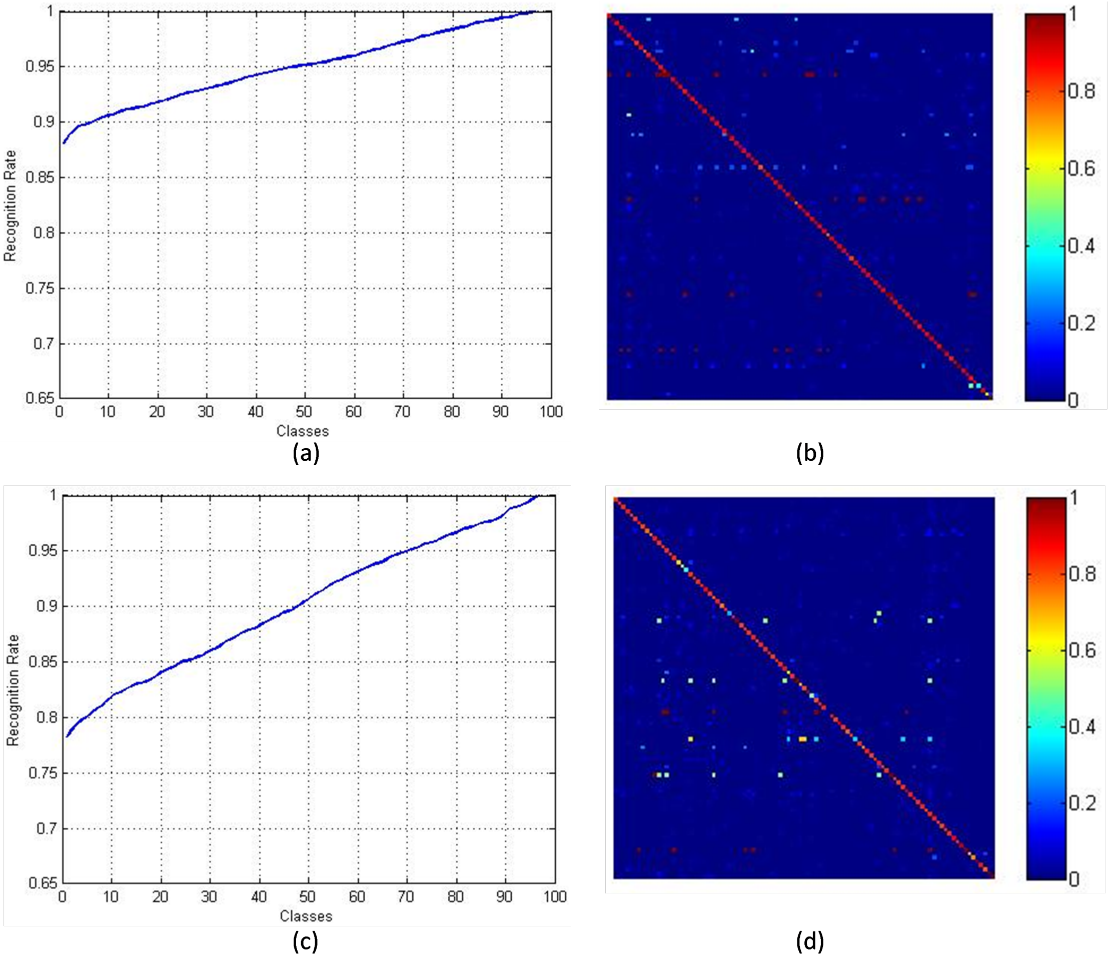


Figure 6.13: Recognition rates and confusion matrices on video datasets. (a) Rank-wise recognition rate on High Quality Video dataset. (b) Confusion Matrix showing classification accuracy across each class in High Quality Video dataset.(c) Rank-wise recognition rate on CCTV Video dataset. (d) Confusion matrix showing classification accuracy of proposed framework on the CCTV Video dataset.

requires an accurate image correction without any effect of in-plane rotation to perform well.

6.4.7 Implementation details and Processsing Time

In order to implement the various stages of the framework, a very effective and efficient publicly available computer vision algorithm library VLFeat [160] is used. The optimized functions for DoG based stable key-point detection, patch extraction, SIFT computation and FV encoding are available in the library. The package routine to handle the large number of patches for feature descriptor computation, mid-level representation while training and testing have been implemented in MATLAB. The proposed system takes an average 3.7 seconds to compute the mid-level representation for an RoI image of 128 by 256 pixels.

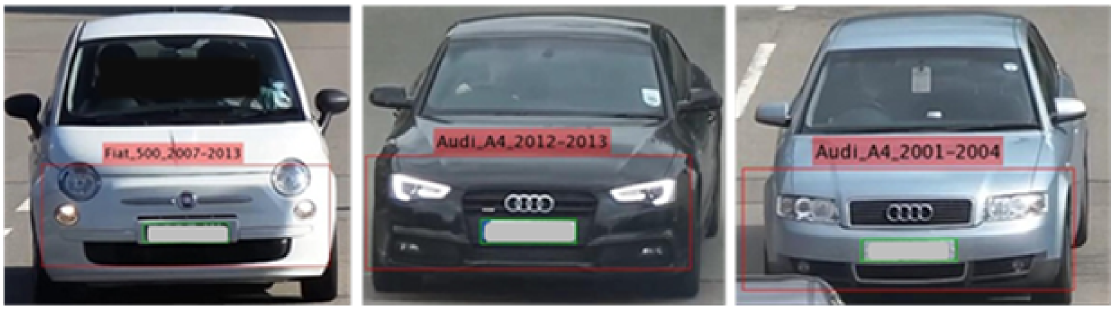


Figure 6.14: Correct recognition on High Quality Video Dataset.

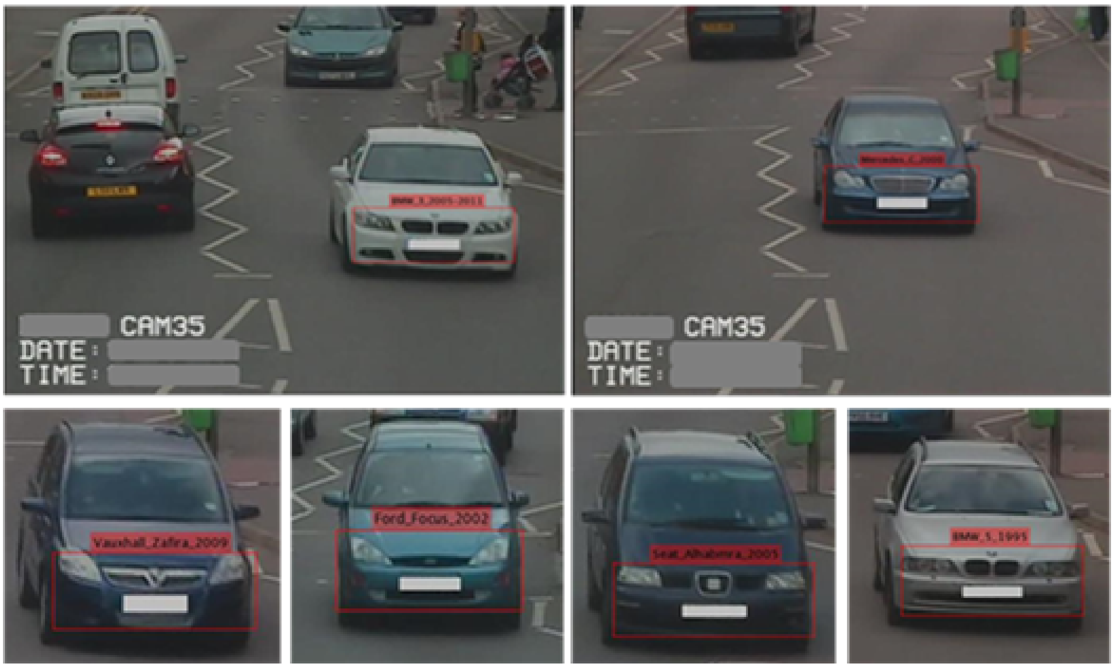


Figure 6.15: CCTV Frames showing correct recognition.

6.5 Summary and Conclusion

In this chapter, a novel visual word based pattern representation technique has been proposed witch is effectively utilized in VMMR. It has been illustrated with extensive experimentation on a number of datasets that an image representation using MLR based visual words captures fine grain discrimination. The generalization ability of the learned model was demonstrated by cross dataset testing. The framework achieved state-of-the-art performance on an existing dataset and three new datasets collected by the author that include up to date car makes and models. A significant benefit of the proposed approach is that it does not need retraining for new vehicle models because of the strong generalization ability of the learned model.

It is worth noticing that all existing techniques rely on conventional global feature representation and key-points based matching systems. These techniques perform reasonably well. However, the chances of their failure increase with the

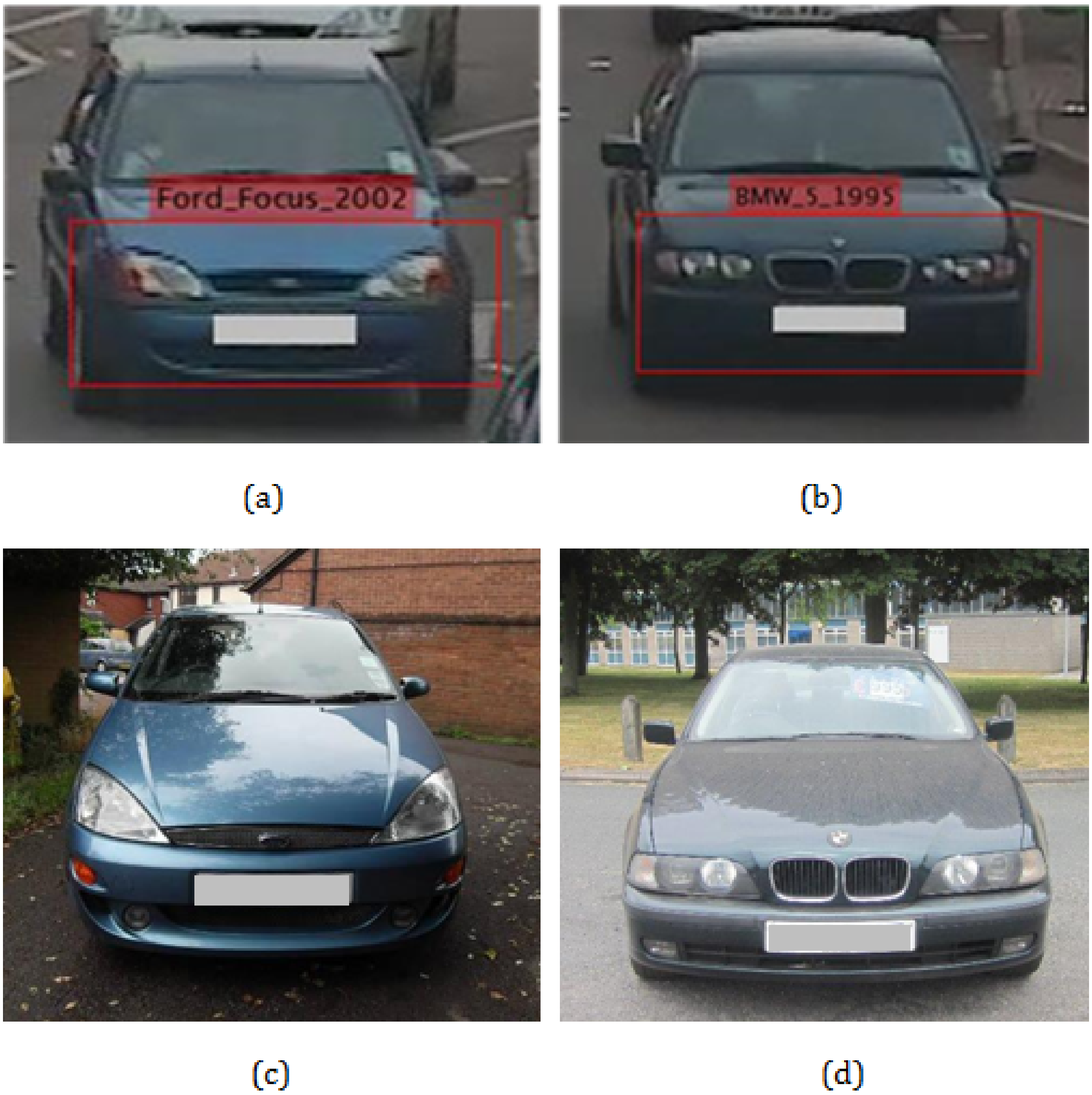


Figure 6.16: A couple of examples of wrong model recognition. (a) Ford-Fiesta-2001 is classified as Ford-Focus-2002. (b) BMW-3-1998 is classified as BMW-5-1995. (c) Image of a Ford-Focus-2002. (d) Image of a BMW-5-1995.

increase in noise and inconsistent illumination within the image. The use of key-points detection eliminates these chances while the construction of MLR based lexicon efficiently aggregates the feature vectors and captures the most distinctive information associated with a representative class. The flexibility that this approach offers in terms of RoI selection is also significant. It is not trivial to devise a universally applicable criterion of RoI extraction because of varying sizes and designs of the vehicles. An inaccurate RoI extraction results in clipping of important information associated with the vehicle in some cases or in other case it takes in unnecessary information from the background region that affects the overall performance. Conclusively, a highly accurate and robust framework has been proposed for the task of vehicle make and model recognition.

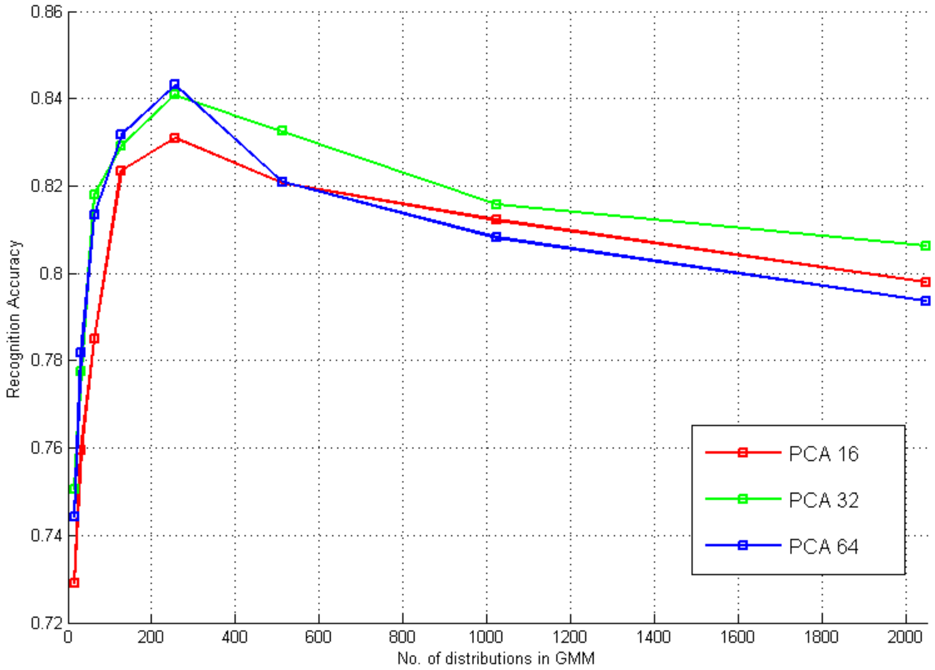


Figure 6.17: Recognition Accuracy with different values of PCA and GMM distributions.

Chapter 7

Detection and Recognition of Text using Colour Information

7.1 Introduction

Text detection and recognition are two major but interlinked problems with numerous applications. Text serves as an important parameter for image and video analysis. It is used for indexing and understanding of images and videos for text based search. A variety of methods have been recently proposed that explore various theoretical and practical aspects of the field to solve the problems of text detection and recognition. The techniques presented in this chapter are based on references [47] and [46]. In this work, particular attention has been paid to both problems individually in designing a framework that detects and recognizes the text in outdoor scenes. The proposed framework makes efficient use of colour information to detect text regions in images. In word recognition scenario, it accurately segments and classifies the characters to recognize words in the images.

Text carries an important characteristic, that is, its colour in comparison to its background. It stands out to its background at least to an extent where humans can identify it. Another important trait, that is, the characters in a word usually possess similar colour that helps the reader to identify the letters of a word. From the detailed review of literature in chapter 2 (section 2.5), it is noticed that none of the existing techniques have so far exploited the colour information up to full potential in devising an end-to-end text detection and recognition system.

The work in this chapter is presented to fill the gap in the literature by intelligently exploiting the above mentioned colour characteristics of text to solve the problem of text detection and recognition in outdoor scenes. First, text detection is performed that involves the localization of words in the images. Next, the characters in words are identified and recognized. Finally, the recognition of words is performed by verifying the combination of recognized characters against a lexicon to remove errors that occur in the character recognition stage.

For detection task, a novel Connected Components (CC) based strategy has been proposed that uses colour information present in the image to extract candidate regions. All existing CC based techniques rely on the edge or gradient information to extract candidate regions due to which they fail in scenarios such as where the image is blurred, text regions have low contrast with their surroundings and the intensity information is non-uniform due to surrounding light. To deal with these problems for candidate region extraction, the input image is preprocessed using colour constancy technique (mentioned in chapter 5) to reduce the effect of illumination and surrounding reflections. The colour corrected image is passed through a noise reduction stage where the colour information in small areas is enhanced while maintaining a sharp boundary between the regions of different colours. It is done in such a way that all the pixels in significant colour areas maintain the same colour. The enhanced image is fed into a colour quantization stage and the CCs in the binary map of each quantization level are used to extract candidate regions. The candidate regions are then classified into text and non-text regions using shape features and a pre-trained classifier.

For the recognition task, the region grouping method and the object recognition strategy is combined to achieve the advantages of both techniques. First, the word image is binarized using colour information and foreground segmentation is performed using modified Bilateral Regression (BR) to separate characters from the background. After that, shape features on binary images of characters are extracted and classification is performed using a pre-trained classifier. The combinations of recognized characters are fed into a string similarity matching stage where lexicon based search is performed to find the closest matching word.

The proposed character recognition pipeline outperforms the current state-of-the-art techniques by a significant margin on Chars74k [18] and ICDAR03-Character [147] benchmark datasets. Further to that, the proposed word recognition pipeline outperforms state-of-the-art approaches on challenging ICDAR03-Word [147] and SVT [168] benchmarks.

The rest of this chapter is organized as follows: In section 7.2 the proposed framework for the detection of text is described in detail. Section 7.3 covers the details about contribution towards text recognition in word images. The experimental evaluation of the proposed method and discussion about results is given in section 7.4. Section 7.5 summarizes and concludes the work proposed in this chapter.

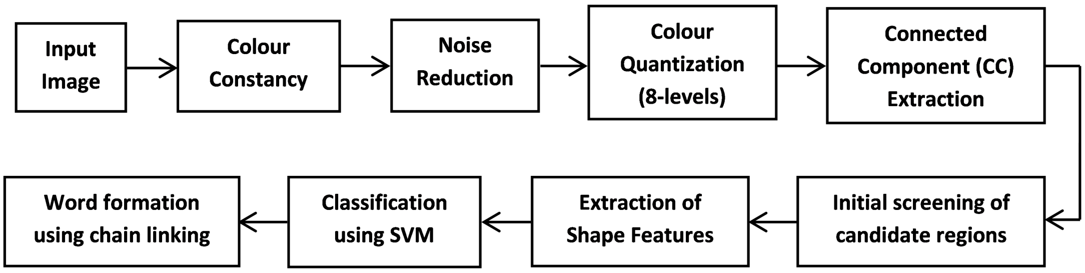

Input Image
Colour Constancy
Noise Reduction
Colour Quantization (8-levels)
Connected Component (CC) Extraction
Initial screening of candidate regions
Extraction of Shape Features
Classification using SVM
Word formation using chain linking

Figure 7.1: Block diagram showing key stages of text detection and pipeline.

7.2 Text Detection

The important stages of proposed text detection framework are shown in figure 7.1. All the stages are explained separately as a sub-section below.

7.2.1 Extraction of Colour Based CCs

The colour based CCs generation initiates with an enhancement stage that improves the colour in the input image. Since the target is to extract candidate regions using colour difference of text from its surrounding therefore it is important to eliminate the effect of surrounding illumination. The colour enhancement aims at reducing the effect of illumination and surrounding reflection using colour constancy mechanism. The Gray World (GW) [180] algorithm is selected in this work for illuminant estimation based on the reason explained in chapter 5 (see section 5.2.1).

7.2.2 Noise Reduction

After colour constancy, the output image is passed through a noise reduction stage. This stage aims at normalizing the pixels of similar colour while maintaining a sufficient contrast between background and foreground pixels by preserving the information at edges. The noise reduction is important as it reduces the inconsistencies between the pixels of the similar colour and resulting in compact pixel clusters at colour quantization stage.

As explained in chapter 5 (section 5.2.2) that the most commonly used Averaging Filter does not preserves edges. On the other hand, the specialized edge preserved image smoothing techniques (BF [152], Guided Filtering [55] and Anisotropic diffusion [113]) are computationally expensive. For instance, Guided Filter is the fastest among these techniques yet it takes 2.75 seconds on average to process an image of 800-by-1200 pixels. This results in a considerably slow detection process. To find the middle ground between two requirements (edge preserving and fast computation), the Median Filtering (MF) [60] technique is used here as it offers the edge preserving capability for small and medium level of Gaussian noise and is computationally faster than specialized edge preserving methods.

It is observed that the choice of kernel size plays an important role here as the features in the image smaller than the kernel size are filtered out while retaining the relatively sharp edges between features in the image larger than the aperture size. Further, the corners, unlike edges, are not well-preserved by the MF and they tend to be blurred to a degree proportional to the size of the median. Hence, the kernel size for the MF is empirically selected to 5-by-5 pixels to maintain a balance between advantages and weaknesses it offers.

7.2.3 Colour Clustering

In order to acquire the clusters of pixels having similar colour, the enhanced image is fed into a colour quantization stage where the full range of colours is reduced to N. For this purpose, the most commonly used k-means clustering technique [86] has several disadvantages in this scenario. First, it is computationally expensive and requires a fairly long convergence time particularly in images where the pixel values are distributed in non-globular manner. Second, the k-means algorithm randomly selects the initial partitions that may result in different final clusters on the same data in various iterations. On the other hand, the Uniform Quantization [86] offers extremely fast reduction of colour levels in the image but it is not suitable here because of its non-dynamic nature.

Heckbert [56] proposed an efficient and extremely fast non-linear quantization technique named as Minimum-Variance Quantization (MVQ). The minimum variance quantization cuts the colour cube in red, green and blue directions until a pre-specified number of non-empty regions are obtained; it then uses the average colour in each region to create the new reduced palette in an iterative manner. The pixels are clustered together on the basis of the variance between their values. For example, a set of red pixels may be grouped together because they possess small variance from the centre pixel of the group. In this work, the MVQ is used due to its dynamic capability and better computation performance as compared to k-means clustering. The quantization process merges all the noisy areas together at different quantization levels while separately clusters the pixels of text regions from surrounding background pixels. The process is fast and accumulates similar coloured regions with good accuracy even in the presence of noise. The binary map of each quantization level is computed by assigning the value 1 to the pixels that fall into that quantization level while setting the values of the remaining pixels to 0. A total of N binary maps are attained for N quantization levels. The value of N is important as it has the direct impact on the quality of candidate regions. A

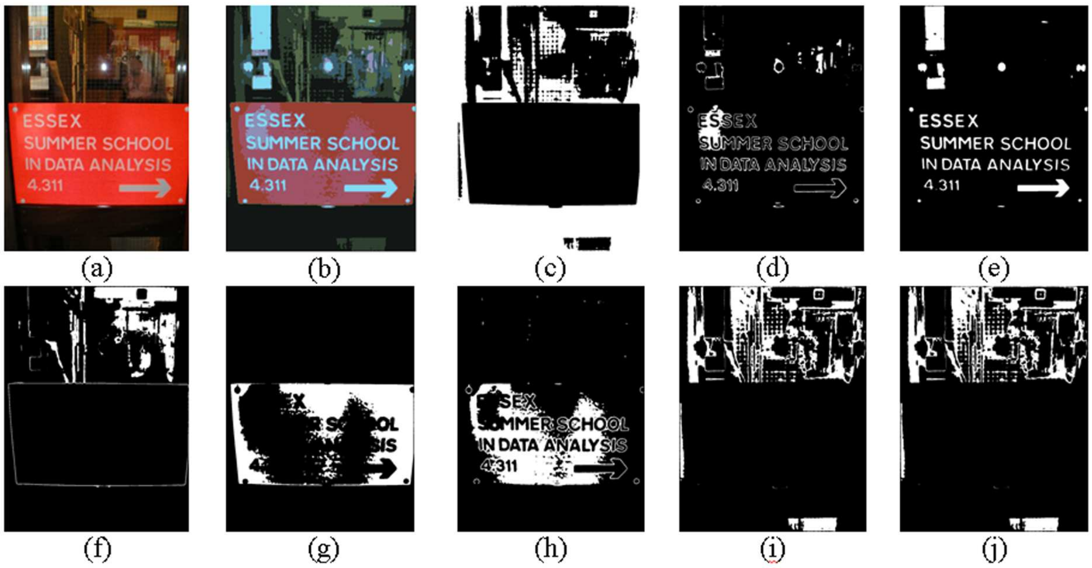


Figure 7.2: Quantization of input image and binary map of each quantization level. (a) Original Image. (b) 8-level quantized image. (c)-(j) The binary maps corresponding to each quantization level. The pixels clustered in each quantization level are represented with 1 while remaining pixels are represented with 0.

small value of N may result in the text regions to merge with background pixels whereas a considerably large value of N may result in the text strokes to break into different quantization levels. The value $(N = 8)$ has been selected empirically in this work because scene images contain a large variety of colours in the background. Figure 7.2 shows the output of the quantization and binarization stage on an image for $N = 8$. Notice, the text is separated from background in a single binary maps (see figure 7.2e) while the background pixels are separately clustered in other binary maps. The background regions are eliminated in the classification stage explained below.

7.2.4 Extraction of candidate regions

The CCs in the binary map of each quantization level are extracted. These CCs are passed through an initial screening where the regions are analysed on the basis of geometric properties (width, height, aspect ratio). The regions that do not fulfil the geometric thresholds are eliminated. The empirically computed upper thresholds th_{u1}, th_{u2} and lower threshold th_l can be mathematically expressed as: $th_{u1} = \frac{2}{3} * im_h$, $th_{u2} = \frac{2}{3} * im_w$ and $th_l = 0.025 * (im_h * im_w)$, where im_h, im_w is the height and width respectively of the input image in terms of pixels. The coordinates of the bounding boxes of the remaining CCs are used to crop respective regions from the original (grayscale) image as the candidate regions. These candidate regions are fed into the feature extraction and classification stage where the regions are classified into text and non-text regions.

7.2.5 Feature Extraction

Each candidate window in resized to n-by-n pixels prior to the extraction of features. This is important because the classifier has been trained using the same sized training examples. The value ($n = 32$) is chosen empirically in this work as it provides a reasonable middle point between the large and the small candidate windows. Next, the Histogram of Oriented Gradients (HOG) feature descriptor for each candidate image is computed. Each 32-by-32 image is divided into 16 sub-blocks of 8-by-8 pixels and HOG features are computed on each sub-block. Finally, the histograms from all the sub-blocks are concatenated to form a single HOG vector. The images are passed through a DoG filter prior to HOG computation to enhance edges and other details that may be affected due to resizing operation.

7.2.6 Classification

A Support Vector Machine (SVM) classifier is trained to discriminate text windows from non-text windows. The training images were extracted from ICDAR11 dataset using the colour based processing and candidate region extraction scheme explained in the preceding paragraphs. A total of 10000 images were extracted for training (5000 positive and 5000 negative). The feature vectors of the training images and the ground truth “text” and “non-text” labels are then used to train the SVM classifier. The soft margin parameter C and Radial Basis Function (RBF)kernel’s γ for the SVM model is computed through 5-fold cross validation on the training data.

7.2.7 Word Formation

The final step in the detection process is the formation of words from detected text windows. In most images, the characters are detected separately because each character forms a single connected component and is separately extracted as a text region. The word formation step is important for performance comparison on benchmark datasets as the ground truths in these datasets is available in the form of word windows. The detected text regions are analysed using colour and location cues. The text present in a linear alignment is expected to have similarities in terms of colour and size. These minor cues serve as a valuable reference to combine letters into chains.

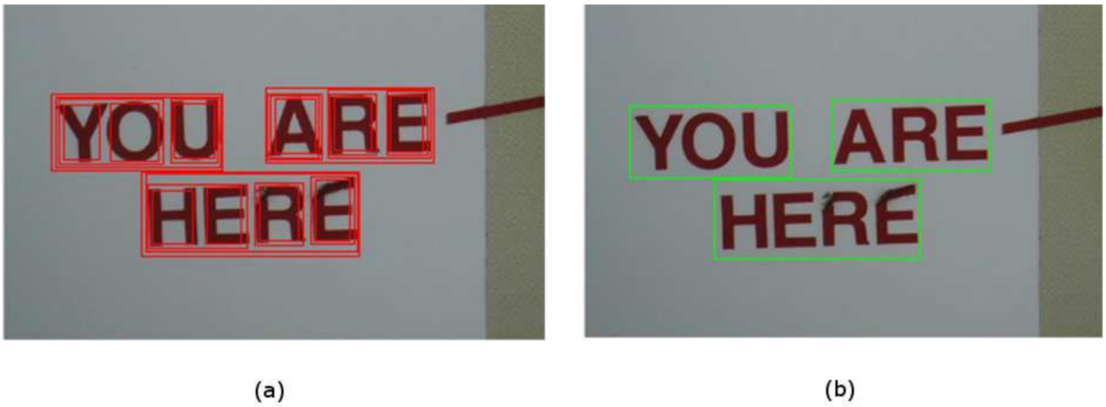


Figure 7.3: Word formation from detected text regions. (a) Windows classified as text regions. (b) Word windows after merging the adjacent detected text windows with similar characteristics.

All the detected text regions are analysed with respect to their adjacent text regions. If two nearby regions possess similarities in colour and size then the bounding boxes of these text regions are merged into a single bounding box. The process is repeated in an iterative manner until all the similar regions are merged into single bounding box. However, there is a chance that two adjacent words merge into a single word. To deal with this, a post processing step is added where the words are separated from each other on the basis of horizontal distance. The geometric distance between the letters of each word is computed using horizontal projection of that bounding box. If the distance between letters is found to be greater than a particular threshold then it is an indication for breaking the chain into two words and so on. The threshold for separating the letters is empirically computed using the window sizes. The threshold varies for different sentences and images. Figure 7.3 shows the detected regions before and after the word formation stage.

7.3 Text Recognition

The next stage after detecting a text region is the recognition of information present in that word. The key stages of proposed text recognition framework are shown in figure 7.4. All the stages are explained separately as a sub-section below.

7.3.1 Character Identification

The text recognition stage starts with character identification. The proposed text recognition module deals with the recognition of text in cropped words. The reason for involving this step is to enable the system to work as a standalone word

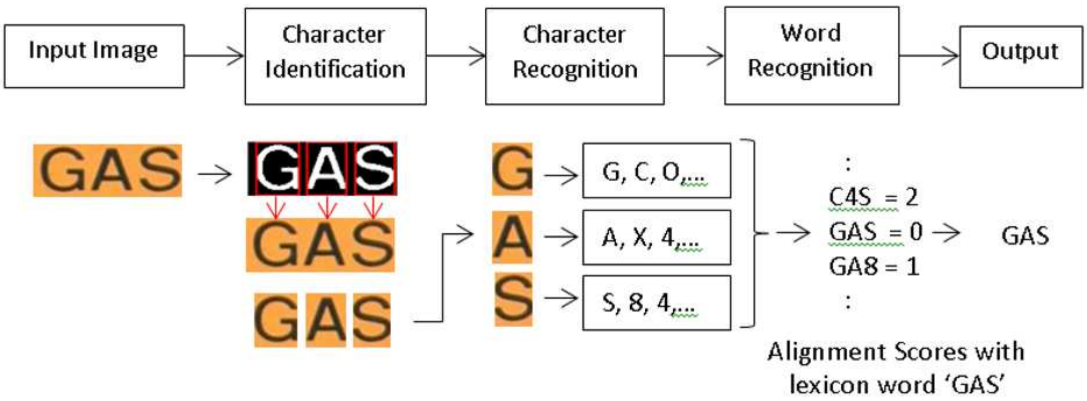


Figure 7.4: Block diagram showing important stages of word recognition framework.

recognition system. The key requirement for character identification framework is the accurate segmentation of characters from background in such a fine way that even the closely located characters remain separated from each other. The gradient or edge based connected components extraction methods reviewed in chapter 2 do not perform satisfactorily for character identification. Most of the recent methods use a sliding window based approach for character identification. This however generates a large number of candidate regions that requires extensive evaluations. Following [42], the BR based technique is used here to segment the characters. However, in this work the proposed approach is different than the original method in such a way that it has been applied only to estimate the horizontal location of each character in the word image. The objective here is the estimation of the starting column and width of each character in word image.

The BR technique models the foreground pixels by using a weighted regression that assigns weight to each pixel according to its location with respect to the foreground in the feature space. The pixels of foreground region get higher weights in comparison to the pixels present in background region. In this case, the regression model in eq. 7.1 represents the quadratic surface that best models the image as a function of pixels location.

$$z = ax^2 + by^2 + cxy + dx + ey + f \tag{7.1}$$

The error between each pixel in the image and the model is computed. The pixels with error value higher than a particular threshold are excluded as background while the remaining pixels are labelled as foreground pixels.

The BR models top n colours in each image separately. In a post processing procedure it chooses the segmentation that is most likely to contain the foreground text by comparing the shape descriptors of foreground regions with a training set. This makes the overall process complex and leads to false segmentation results.

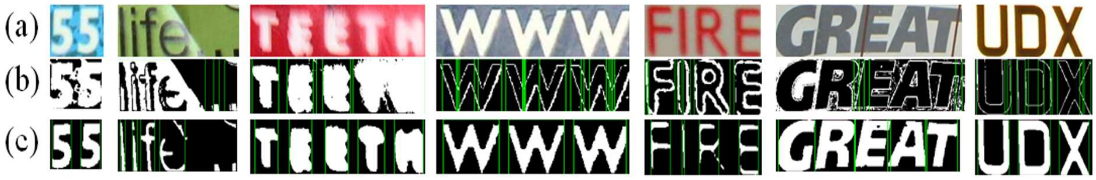


Figure 7.5: Improved character identification in word images. (a) Original Images. (b) Character segmentation using BR. (c) Character segmentation using proposed pre-processing and BR.

The operation of bilateral regression has been enhanced in the proposed system by incorporating a pre-processing step where the foreground colour is estimated, a priori. An N-level colour quantization is applied to the word image and binary map of each quantization level is computed. Here, keeping in view the relatively smaller variance of colours in the cropped word as opposed to the whole scene image, each word image is quantized into three colours ($N = 3$). The main motivation is the observation that in the word image the background (having the uniform colour) can be captured in one large colour cluster while most of the foreground character pixels are captured in another cluster. The small variations typically present along the edges of the characters due to noise and illuminations can be captured in another small cluster. The respective binary map for each quantization level is separately analysed to decide about the foreground binary map.

The binary map that contains the highest number of white pixels along the border is classified as the background binary map and is dropped straight away. From the remaining two binary maps, the total number of white pixels present in each binary map are counted. The binary map that contains less white pixels is dropped and the third binary map is characterized as the foreground map. The average colour value of the pixels belonging to the foreground region is then used to perform BR.

The characters are cropped from the original (coloured) word image using the estimated horizontal location while the height is kept same as the height of the original word image. In this way, some background information might also be included but the chances of segmenting only a part of character are reduced. The segmented characters from original word image are fed into the character recognition pipeline explained next. Figure 7.5 depicts the improvement achieved in character identification using the proposed pre-processing procedure in BR.

The segmented masks are used to crop the characters from original (coloured) image and fed into the character recognition pipeline explained in the next section.

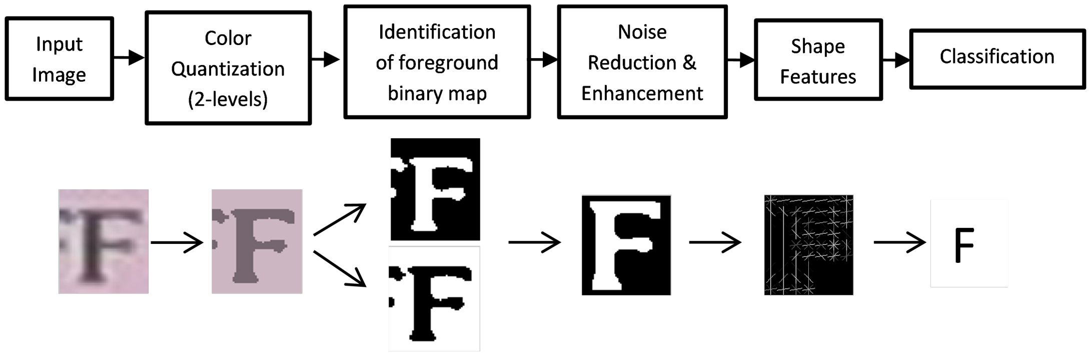

Input Image
Color Quantization (2-levels)
Identification of foreground binary map
Noise Reduction & Enhancement
Shape Features
Classification
F

Figure 7.6: The block diagram of the character recognition framework and the visualization of output at each stage.

7.3.2 Character Recognition

Accurate character recognition has a significant importance in achieving correct word recognition. After a character is properly segmented the recognition requires to be robust with regards to the noise, illumination variations and perspective distortions. Figure 7.6 shows all of the important steps of character recognition. This following section describes each step in detail.

Foreground Segmentation: Consider an image containing a character along with background noise. Similar to the character identification stage, MVQ is applied to enhance the character. Note that the foreground binary mask obtained for each character in the previous stage cannot be directly used. Although it gives a good separation in terms of segmenting the character from the detected text the character itself may not be well represented because of the missing edge pixels. It was empirically observed that for a character image, a 2-level ($N = 2$) quantization is sufficiently good to recover the full character pixels from background.

Next, two binary maps corresponding to the two colour levels are generated by assigning the pixels of each colour cluster a value of 1 (white). Similar as previous stage, one of the two binary maps is categorized as foreground character map. This is done by computing the density of white pixels along the borders of each binary map. The binary map of the background tends to have more white pixels along the borders as compared to the binary map of the foreground. The foreground binary map comprises of mostly the pixels associated with the character region therefore it possesses very low density of white pixel around the borders especially in the corners. This property is exploited here to classify the binary maps as foreground and background. A 5-by-5 pixel window in each of the four corners of the binary map is selected and the total number of corner white pixels in each binary map is counted. The binary map that possesses the higher total number of corner white pixels is considered as the background and the other binary map is classified as

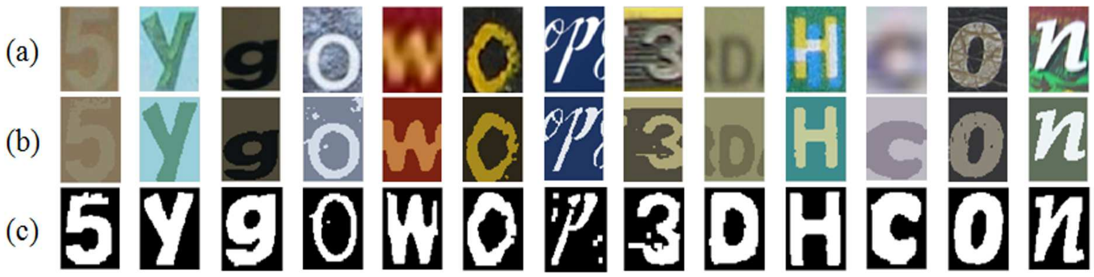


Figure 7.7: Examples of character images from ICDAR03-CH and Chars74k datasets where the proposed method extracts and enhances the binary map of the characters. Row (a) shows the Input Images, row (b) represents 2-level minimum variance quantization of input images and row (c) shows the enhanced binary maps of the characters with reduced noise.

the character map.

Noise Reduction and Enhancement: The foreground binary map is passed through an enhancement and noise reduction pipeline where the unwanted pixels are eliminated. Morphological closing, spur removal and dilation operations are applied to remove noisy pixels and enhancement of character pixels. The CCs are computed in the foreground binary map and the biggest CC is selected as the character. The binary map of the character is resized to k-by-m pixels and padded with an array of five black pixels on all sides resulting in a binary image of size (k+10)-by-(m+10) pixels with character map perfectly centred in it. The value of k and m is empirically selected as 64 and 48 respectively in this work. It is observed that the characters are slightly taller than their width and the selected size approximately maintains the original aspect ratio of the characters. Figure 7.7 shows a few images where the proposed colour based binarization scheme accurately extracts and enhances the character binary maps in challenging scenarios.

Features Extraction and Classification: The binary map acquired in the previous step is directly used for feature extraction. HOG features are computed to capture the shapes of the characters for classification. The HOG features are computed directly on the binary maps of the characters extracted in the previous stage. For classification, a multi-class SVM is used. In this work, the character classes are divided as the Digits (10 classes) and English letters (52 Classes) i.e. the alphabets ζ = 0,...,9;A,...,Z;a,...,z and $|\zeta|$=62. Hence, a 62-class non-linear SVM is trained in a one-vs-all manner. The best parameters for training the SVM model have been estimated using Radial Basis Function (RBF) kernel and 5-fold cross validation.

7.3.3 Word Recognition

The word recognition stage requires the accurate recognition of characters identified in a cropped word image. The inaccuracies in character identification and recognition may easily lead to false alarms. Since, the improvements have been proposed in both prior stages; the recognition framework relies on a simple lexicon based alignment procedure to remove such errors that occur in character identification and recognition. Some errors in character recognition are inevitable because of high interclass similarity between various characters i.e. 'I', '1' and 'l', '0', 'o' and 'O' etc. The alignment procedure is performed using Lavenshtien distance [98].

Alignment with Lexicon: The character recognition pipeline predicts the character label l on the basis of highest probability estimate. In order to deal with wrong recognitions the first predicted label cannot be relied upon especially if the probability estimates for that recognition is too low. To deal with this, top η predicted labels are selected based on a confidence score S_c. The confidence score is the sum of probability of top η predicted labels. The value of η varies in every test case. When S_c approaches a threshold τ, the character labels corresponding to those probability estimates are included in the predicted word combination. The value $\tau = 0.35$ has been experimentally selected in this work.

One potential problem that may occur in this set up is the formation of a large number of character combinations for certain images. In order to avoid that, a limit is put on the maximum number of words. A total of 30 words that possess the highest sum of probability estimates of characters are selected. The procedure is precisely expressed in the algorithm 1.

```
for i= 1:len(word) do
    $\eta$=1;
    Assign $S_c$ = P($\eta$) ;          // P($\eta$) is the Probability of the $\eta$th
    predicted label.
    while ($S_c$ ¡ $\tau$ and $\eta \leq$ 30) do
        EstimatedWord($\eta$,i) = l($\eta$) ;    // Update word combinations by
        including $\eta$th label l($\eta$).
        $S_c$ = $S_c$+P($\eta$) ;    // update confidence score by adding next
        highest probability.
        $\eta$=$\eta$+1;
    end
end
```

Algorithm 1: Computing words using the combination of the characters with high recognition probabilities.

Next, the correct word is searched from all the predicted words. This is done by

$$lev_{s_1,s_2}(i,j) = \begin{cases} \max(i,j) & if\, \min(i,j) = 0 \\ \min \left\{ \begin{array}{c} lev_{s_1,s_2}(i-1,j)+1 \\ lev_{s_1,s_2}(i,j-1)+1 \\ lev_{s_1,s_2}(i-1,j-1)+1_{(s_{1i} \neq s_{2j})} \end{array} \right\} & otherwise \end{cases} \tag{7.2}$$

the alignment of predicted words with a pre-stored lexicon using a string similarity measure. The closest matching word in the lexicon is finalized as the word present in the image. The Lavenshtein distance between two strings is used to compare the predicted word string with the words in the lexicon. The Lavenshtein distance computes the total number of operations (insertion, deletion, and replacement) required to align a string with the other. Mathematically, the Lavenshtien distance between two strings s_1 and s_2 can be computed using equation 7.2.

The accuracy of word recognition framework directly depends upon the performance of the character recognition framework. The proposed character recognition framework (in sec. 7.3.2) works exceptionally well as a result of which the simple lexicon alignment strategy achieves accurate word recognition.

7.4 Experiments and Results

The experimental setup and results for each stage of the proposed framework are discussed in detail in each subsection.

7.4.1 Text Detection

The proposed text detection method is evaluated on two benchmark datasets ICDAR11 [134] and extremely challenging Street View Text (SVT) dataset [168].

Performance Evaluation: To evaluate the performance of the proposed text detection pipeline, the standard performance measures (Precision, Recall and F-measure) have been used. Following [174], the object level precision and recall is computed first. In the next step, the harmonic mean (F-measure) is computed by taking the mean value of object measures (Precision and Recall) over all possible constraint values. The values of the constraints while computing these performance measures have been used the same as given in [174].

ICDAR11 datset: The ICDAR11 dataset contains 229 training images and 225 test images with sizes varying between 626 × 179 and 3888 × 2592 pixels. The proposed framework achieves a precision of 80.90% which is in par with existing

Table 7.1: Comparison of the proposed text detection method with other techniques on ICDAR11 dataset.

Method	Precision	Recall	F-measure
Huang et al. [61]	0.82	0.75	0.73
Neumann and Matas [102]	0.82	0.75	0.73
Shi et al. [138]	0.83	0.63	0.72
Proposed Method	**0.81**	**0.64**	**0.71**
Kim et al.[134]	0.83	0.62	0.71
Neumann [101]	0.73	0.65	0.69
Yi [184]	0.76	0.68	0.67
Gonzalez [50]	0.73	0.56	0.63
Yi [183]	0.67	0.58	0.62
Neumann [100]	0.69	0.53	0.60

Table 7.2: Comparison of proposed text detection method with other techniques on SVT dataset.

Method	Recall
Proposed Method	**0.52**
Neumann [101]	0.33
Wang et al. [168]	0.29

state-of-the-art; however, the recall rate of 63.57% is slightly poor resulting in the f-measure score of 71.26% which is marginally low as compared to the existing start-of-the-art. This is justifiable due to the fact that proposed technique is simple and computationally less expensive as explained in section 7.4.5. Table 7.1 presents the performance comparison of the proposed method with other techniques on ICDAR11 dataset.

SVT Dataset: The SVT dataset is collected from Google Street View and comprises of 101 training and 249 test images. The dataset is extremely challenging because the images contain significant background information with similar pattern as text regions. Also, the words are not in fully horizontal orientation. Another challenge is that the dataset is not fully annotated. This is the reason only a few authors in the literature have provided the text detection results on this dataset.

The proposed method has achieved a recall rate of 51.55% and superseded the state-of-the-art techniques by a considerable margin. Table 7.2 presents the performance comparison of proposed method with other techniques on SVT dataset. Since that dataset is not fully annotated therefore the precision and F-measure scores are not meaningful and hence the methods have been compared only on the basis of recall score.

7.4.2 Character Recognition

To evaluate the performance of the proposed character recognition method, two well-known benchmark datasets for scene character classification task have been used. These are:

1. Chars74K-15 dataset

2. ICDAR03-CH dataset

Chars74K dataset contains 12505 images of characters extracted from scene images. The images are divided into 62 classes (52 alphabets and 10 number digits). The authors have provided various training and test splits of dataset for the research community to perform fair comparison of results. Chars74K-15 is one subset that contains 930 (15 per class) training images and another 930 (15 per class) test images from Chars74K dataset.

ICDAR03-CH dataset contains 11482 images of characters. This does not include non-alphanumeric characters. The dataset comprises of 6113 training and 5369 test images.

Table 7.3 lists the character classification results. The proposed framework superseded other techniques by achieving the classification accuracy of 81.34% on ICDAR03-CH dataset and 72.03% on Chars74K-15 dataset. It is evident from results that the non-informative regions around the characters have significant degradation impact on recognition accuracy. Also, the intensity variation within the strokes of characters causes a negative impact on recognition accuracy. The proposed framework not only removes the noisy background regions to achieve a nicely centred binary map but also attenuate the effect of intensity variation within the strokes of characters. This results in a strongly discriminative shape feature extraction along the contour of the characters.

In order to evaluate the importance of noise reduction and enhancement stage, the experiments have been performed where the character images are binarized without applying any noise reduction and enhancement procedure. The remaining character recognition pipeline is followed as explained in section 7.3.2. Table 7.4 shows the character recognition performance with and without applying the proposed noise reduction and enhancement stage. Notice the significant performance degradation in the absence of noise reduction and enhancement stage.

The proposed methods in [18, 77, 168] merge similar character classes from 62 to 49 to reduce the confusion among similar looking characters for example ’0’, ’O’, ’o’ and ’1’,’l’,’I’ etc. For a comparison, the proposed character recognition pipeline

Table 7.3: Comparison of the proposed scene character recognition accuracy (%) with other techniques.

Method	ICDAR03-CH	Char74K-15
Proposed Method	**81.34**	**72.03**
MLFP [77]	79	64
GHOG+SVM [185]	76	62
LHOG+SVM [185]	75	58
HOG+NN [168]	52	58
MKL [18]	-	55
NATIVE+FERNS [168]	64	54
GB+SVM [18]	-	53
GB+NN [18]	-	47
SYNTH+FERNS [168]	52	47
SC+SVM [18]	-	35
SC+NN [18]	-	34
ABBYY [18]	21	31

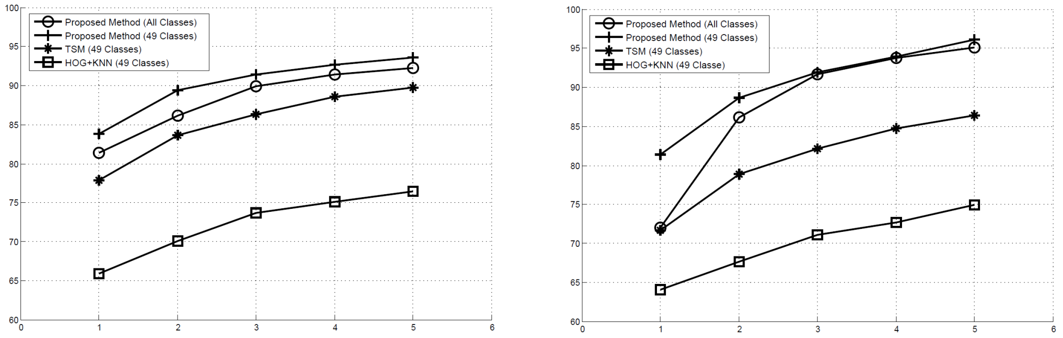
Table 7.4: Character classification performance with and without the use of noise reduction and enhancement stage.

	ICDAR03-CH	Char74K-15
With noise reduction stage	81.34	72.03
Without noise reduction stage	74.87	58.60

is also evaluated in 49-classes setup. Table 7.5 compares the character classification performance in 49-classes setup where the proposed framework outperforms the other techniques with a substantial margin especially on Chars74k-15 dataset. Figure 7.8 shows the rank-wise scores for top 5 probabilities of character recognition. The recognition rates for top five candidates approach 94% for 49 classes and 92.5% for all (62) classes on ICDAR03-CH dataset which indicates that the use of top 5 character recognitions shall also enhance the word recognition performance.

Table 7.5: Scene character recognition accuracy (%) comparison with the techniques of [18, 77, 168] on ICDAR03 and Chars74k benchmark datasets is 49-classes setup. The proposed method in [77] use discriminative mid-level feature pooling, the technique in [18] use tree-structure based character classifier and the proposed method in [168] use HOG+NN based character classification.

Method	ICDAR03-CH	Char74K-15
Proposed Method	**82.77**	**81.39**
MLFP [77]	81	74
MKL [18]	77.86	71.67
HOG+NN [168]	65.92	64.02



ICDAR03-CH Dataset Chars74k Dataset

Figure 7.8: Rank-Wise character recognition scores on ICDAR03-CH and Chars74k datasets for all (62) classes and for 49 classes. The results have been compared with [77] and [168] where they reported the efficiency of their methods on 49 classes.

7.4.3 Word Recognition

To evaluate the performance of proposed word recognition frame work the benchmark datsets: ICDAR03-WD, ICDAR11-WD and SVT-WD have been used. For a fair comparison the same training and testing splits and the lexicons have been used as in [168]. A number of different sized lexicons have been provided in [168]. This work used ’FULL’ and ’50’ lexicons to evaluate and compare the proposed framework. ’FULL’ lexicon contains all the words from test set of ICDAR03-WD dataset whereas in ’50’ there are 50 distracting words. Notice that this work use ICDAR03-CH data for training. Following [103, 93, 139, 77, 181], the words with 2 or fewer characters as well as those with non-alphanumeric characters have been skipped.

The proposed framework achieves 89.37% and 88.94% word recognition accuracy on ICDAR03-WD and ICDAR11-WD respectively in small (50 words) lexicon setup and outperformed existing techniques. The performance in large lexicon (all words) setup is 77.47% and 78.58% respectively on both datasets which is slightly less than the existing state-of-the-art techniques. The performance degradation indicates the weakness of the word recognition method in the presence of large number of distracting words. The recognition performance on SVT dataset is 78.66% which is third highest recognition accuracy reported to date. Considering the simplicity of the word recognition pipeline, the performance of the proposed method is promising. Table 7.6 compares the cropped word recognition performance of all the techniques that use the same test setup as [168]. Figure 7.9 shows a few examples where the proposed word recognition framework successfully recognizes the word in the image. More results for end-to-end system are shown in Fig. 7.11.

A drawback of the proposed character identification module is its inability

Table 7.6: Comparison of lexicon based cropped word recognition performance of various methods.

Method	ICDAR03 (Full)	ICDAR03 (50)	ICDAR11 (Full)	ICDAR11 (50)	SVT
Proposed Method	77.47	**89.37**	78.58	**88.94**	78.66
PhotoOCR [12]	-	-	-	-	**90.39**
Strokelets [181]	**80.33**	88.48	-	-	75.89
MLFP+PLEX [77]	76	88	77	88	80
MRF [172]	-	-	-	-	78.05
TSM+CRF [139]	79.30	87.44	**82.87**	87.04	73.51
Mishra et al. [93]	67.79	81.78	-	-	73.26
Mishra et al. [94]	-	81.78	-	-	73.26
Novikova et al [103]	-	83	-	-	73
TSM+PLEX [139]	70.47	80.70	74.23	80.25	69.51
Field et al. [42]	67.76	76.53	-	-	-
SYNTH+PLEX [168]	62	76	-	-	57
ABBYY [168]	55	56	-	-	35

Figure 7.9: Some example images from ICDAR11 and SVT dataset where the proposed framework correctly recognized the word.

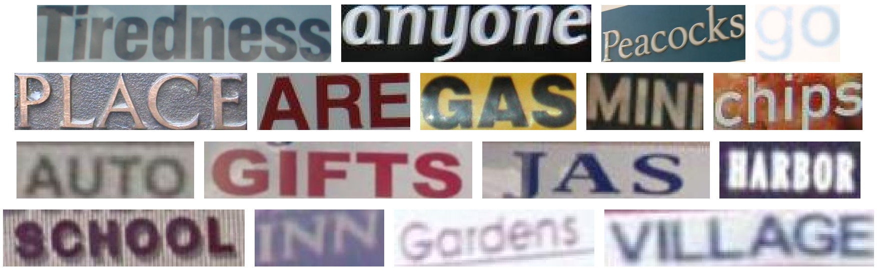


Figure 7.10: A few example images from ICDAR11 and SVT dataset where the proposed framework failed to recognize the word.

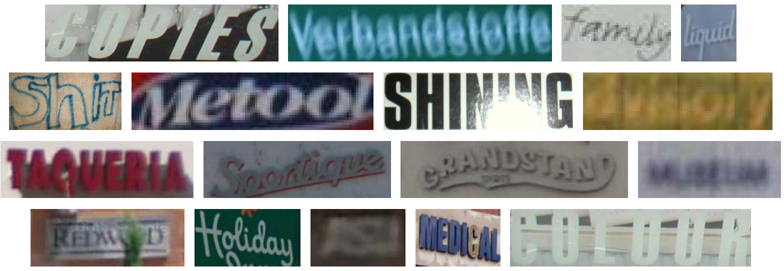


Table 7.7: Recognition accuracy with and without using the proposed pre-processing stage in BR based character identification.

Method	ICDAR03-Full	ICDAR03-50
Proposed modified BR	77.47	89.37
Original BR. [42]	66.19	69.17

to deal with scenarios where the characters are joined either because of font or because of lighting and viewing angle. Figure 7.10 shows a few images where proposed character identification technique failed that resulted in wrong word recognition output. However, the other techniques also fail in such scenario owing to the challenging nature of patterns formed due to illumination and joining of characters.

7.4.4 Evaluation of the Modified Bilateral Regression based Character Identification

The performance of the proposed modified BR based character identification scheme is also evaluated. The experiments have been performed on ICDAR03 dataset using both (50 and Full) sized lexicons. Table 7.7 shows the word recognition performance in two scenarios:

1. When character identification is performed using only the BR based segmentation.
2. When the character identification is performed using the proposed pre-processing stage prior to BR based character identification.

The results clearly depict the significance of the proposed pre-processing stage in BR based character identification.

Table 7.8: Average execution time of each stage of the proposed framework.

	Stage	Time	Total Time
Text Detection	Colour Constancy	0.1 sec	
	Colour Quantization	0.1 sec	
	Candidate extraction and classification	1.2 sec	1.6 seconds
	Word Formation	0.2 sec	
Word Recognition	Character Identification	1.2 sec	
	Character Recognition	0.4	1.7 seconds
	Alignment with Lexicon	0.1 sec	

Table 7.9: Comparison of average processing time of proposed method with other approaches.

Method	Detection (sec)	Recognition (sec)	Total (sec)
Proposed Framework	**1.6**	1.7	**3.3**
Epshtein et. al [40]	6	-	-
Yao et al.[180]	7.2	-	-
Neumann and Matas [102]	-	-	35
Phan et al. [116]	-	38.6	-
Novikova et al. [103]	-	**1.53**	-

7.4.5 Computational Performance

All the components of the proposed framework have been implemented in MATLAB. The average execution time of each module on a standard PC is given in table 7.8. Notice that the code is unoptimized and the execution time can be reduced further with the inclusion of code optimization and parallel processing techniques. Table 7.9 compares the average execution time with other techniques (where available). The average execution time of the end-to-end system per image is considerably low (3.3 seconds) as compared to [102] where the reported execution time is 35 seconds.

7.5 Summary and Conclusion

An end-to-end scene text recognition framework has been presented in this chapter. Experimental evidence showed that the smart exploitation of colour information within the images can generate promising accuracy for the challenging task of text detection and recognition in medium to poor quality CCTV video. A novel pipeline involving colour enhancement operations is proposed to improve the separation of text regions from noisy background. For text recognition, a region grouping based technique (i.e. binarization) is combined with an object recognition strategy to achieve state-of-the-art results on Chars74k and ICDAR03-CH benchmark datasets. It is demonstrated further that instead of using a complex word model, a simple but effective colour exploitation and enhancement procedure leads to better character segmentation and recognition. Based on a string similarity measurement technique, exceptional word recognition accuracy is achieved on ICDAR03-WD, ICDAR11-WD and SVT benchmark datasets. Conclusively, a fairly simple, fast and practically exploitable text recognition scheme has emerged as a result of this work.

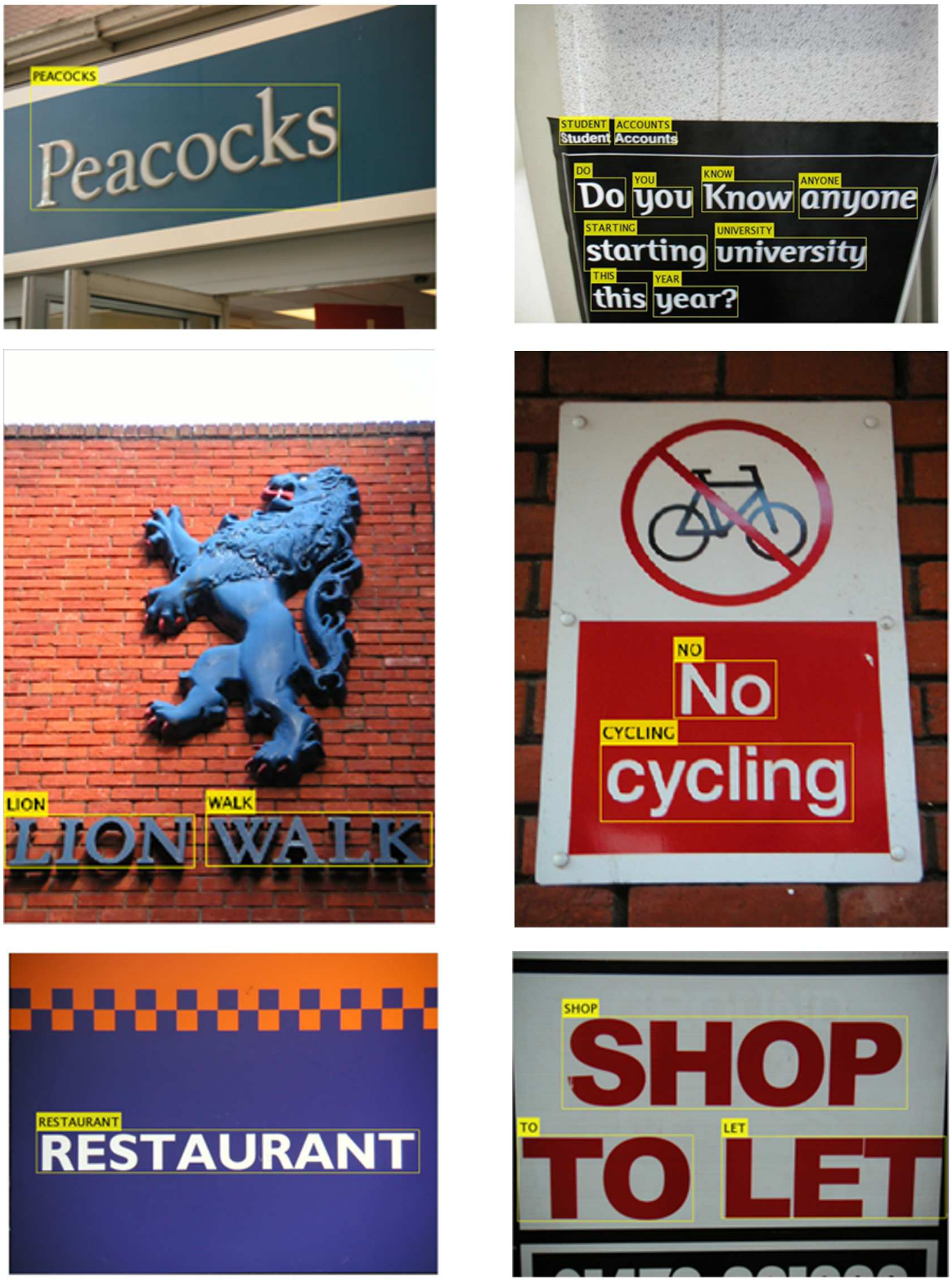


Figure 7.11: Results of end-to-end detection and recognition pipeline.

Chapter 8

Conclusion, Limitations and Future Enhancements

8.1 Introduction

The research in this thesis has focused towards the extraction of various visual parameters for the annotation of surveillance videos for forensic application. The contribution is made to address the problems of moving object identification, colour correction and recognition of objects, vehicle make and model recognition and text detection and recognition. The approaches adopted to address these problems are those that differentiate the current work from other techniques. These approaches, the relevant original contributions made and conclusions are outlined in section 8.2. Section 8.3 discusses the limitations of the proposed algorithms followed by the proposed future enhancements in section 8.4.

8.2 Conclusion

The thesis presents a number of original contributions to the state-of-the-art for the application of video content analysis. The novel Contourlet Transform based Centre Symmetric Local Binary Patterns (CCS-LBP) based feature descriptor proposed in chapter 4 for representing the moving objects in video frames made use of the Contourlet Transform (CT) to enhance the contour segments present in the objects. Due to the fact that the silhouette of the human contains high density of vertically oriented contour segments, it is decided to use only those frequency sub-bands that are achieved from vertically oriented filters. Therefore, the texture in those sub-bands is captured using the LBP. This also reduced the size of the feature descriptor. The selection for the particular variation of the LBP is also made carefully. The CS-LBP technique is used to capture the pattern in the CT sub-bands. It compares only centre symmetric pairs of pixels to produce more compact binary patterns. Apart from the novel feature descriptor the chapter also presented improved foreground extraction technique and an efficient mechanism to separately detect the body parts of the human objects for their detailed annotation.

The colour correction pipeline presented in chapter 5 used colour constancy algorithm and edge preserving BF to remove noise while maintaining a high contrast at the edges. A novel adaptive thresholding procedure is devised to modify the Hue, Saturation and Value components of the image in order to boost the true colours of the objects. In the next phase, the colour recognition is performed for the objects of interest. The proposed colour recognition pipeline exploited MPEG7 Dominanr Colour Descriptor (DCD) to capture the colour information present in the Regions of Interest. A novel mechanism to exploit the temporal information is adopted to compute the colour descriptor of the objects using the pixels collected from multiple frames. The use of temporal information contributes positively towards the performance of the system. Another contribution of the work is the enhancement and accurate recognition of white colour by using the average Saturation and Value information within an object.

In chapter 6, a state-of-the-art technique is proposed to recognize the make and model of the vehicles in images and video frames. A novel visual words based scheme is proposed to capture the texture pattern in the frontal view of the vehicles. The proposed visual word extraction use dense Scale Invariant Feature Transform (SIFT) features and Fisher Vector (FV) encoding to compute the representative visual words for each class of the vehicles. The classification methodology is kept simple by making use of Cosine Similarity (CS) measure. This offers the ease of extending the capability of the system to new vehicle classes without the need of retraining. This capability makes the system more attractive for practical applications. Another important contribution is the collection of up to date vehicle databases. The vehicles make and model recognition work resulted in 3 new benchmark datasets (1 image dataset and 2 video datasets) to aid the research in this domain [available at: http://imaging.lboro.ac.uk/].

Text detection and recognition work presented in chapter 7 efficiently exploited the colour information for isolating the text regions from their background. The proposed method used Minimum Variance Quantization (MVQ) technique to reduce the colours in the image. The connected components in the binary map of each quantization level are analysed separately for their classification as text and non-text regions. Another important contribution of the work is the modified Bilateral Regression (BR) technique for the identification of text pixels in a word image. The technique showed better performance in identifying text pixels as compared to the originally proposed method. The character recognition work exploited the advantage of extracting the shape features directly on the binary images that resulted in a highly distinctive feature representation. This is due to the fact that binary images contain reduced noise and are less affected by the variable colour patterns and texture within the image. The proposed word recognition framework made use of a straightforward lexicon alignment technique that saved a significant amount of computational time. The extensive experimentation on benchmark datasets revealed that the proposed framework is highly accurate while computationally less expensive as compared to the published state-of-the-art.

8.3 Limitations

The proposed algorithms are constrained by several limitations if specific scenarios are to be served. This section critically reviews the underlying reasons for the limitations and the potential impact they may have in practical applications.

To begin with, the method for object recognition strongly relies on the accuracy of the foreground extraction technique. The proposed foreground extraction technique is unable to cater the situations when subjects are moving very close to each other. Similarly, the presence of random patterns in the background is classified as foreground objects and selected as the part of the moving target. The inaccuracy in the foreground extraction causes a negative impact on the feature representation and results in wrong classification of human and vehicle objects. A minor drawback of the proposed CCS-LBP features is the high computational time. The computational time of the proposed method is higher than the baseline technique that uses the combination of Histogram of Oriented Gradients (HOG) and Support Vector Machine (SVM) for human and vehicle detection. The proposed human body parts classification works accurately in most of the scenarios, however, it lacks a more reliable measure especially for the objects that are not in standing position.

The proposed colour correction and enhancement pipeline relies on basic measures to compute the parameters for thresholding the video frame for boosting the true colours. The method uses the mean value in the histogram of the Saturation and Value to compute the parameters for adaptive thresholding. This method may not work well in a few limited scenarios when the histogram of the true coloured image lies beyond the specific threshold. Similarly, a reliable procedure is missing in the current framework for the initial assessment of the video frames and the analyses of the improvements required in a particular scenario. The current framework for colour correction and extraction is dealing with moving objects, so if there are stationary objects in the scene they'll be neglected as background.

The proposed vehicles make and model recognition system is extremely accurate if the frontal view of the vehicle is captured. The system works robustly even with the angular variation of up to 30 degrees but the results start to deteriorate if the angular variation goes beyond that limit. The tests have not been made from the rear view of the vehicle but considering the capability of the proposed framework on frontal view of the vehicles, it is expected that it will perform equally well on the rear view images of the vehicle too. A drawback of the proposed framework is its high computational time and memory requirements. The dimensionality of the visual words is significantly high. The system stores a lexicon of few hundred visual words per class from the training images which results in a significant amount of memory utilization to store the lexicon of visual words from all training images. Similarly, the recognition framework compares a few hundred visual words from the test image with the lexicon of training visual words. This requires a substantial amount of computation time and memory resources.

In the proposed text detection and recognition system, the capability of the proposed character identification technique is limited only to those images where characters are written separately from each other. The proposed method analyses the connected components in the foreground binary maps of the images. The scenarios where characters are joined together, the combination of multiple characters forms a single connected component. Such a region is easy to classify into text and non-text but separating each character using the proposed method is not possible and result in wrong character or word recognition. Similarly, the criterion used for estimating the character binary map only analyses the pixels density along the edges which may sometimes lead to wrong estimation especially when the image is badly affected by noise.

8.4 Future Enhancements

To overcome the above mentioned limitations it is possible to suggest directions for future research.

The proposed CCS-LBP feature descriptor for moving object identification works accurately in almost all the scenarios. However, a dynamic mechanism may be developed for the selection of Contourlet sub-bands at run time. This will not only enhance the robustness of the feature descriptor to varying scenarios but also reduce the feature descriptor size that will eventually put a positive impact on the computational time of the object detection framework. In human body parts identification task, an adaptive method is required to improve the robustness of the method especially to deal with scenarios when human figures do not appear in perfectly standing position.

An important direction for improving the proposed colour extraction framework in chapter 5 is to apply the proposed technique in conjunction with an object recognition strategy. This will not only improve the colour recognition of the objects of interest but also improve the process of object identification. The robustness of the system will also increase with the inclusion of shape information. Similarly in the object retrieval task, the shape information along with the colour of the objects is expected to improve the accuracy of the object retrieval. The shape information regarding the objects have been used to retrieve the respective objects, however, if the DCD is combined with an efficient shape descriptor into a single representation then it can exploit the benefits of both features for efficient retrieval of objects. As mentioned in the section 8.3 that the adaptive thresholding mechanism lacks robustness and reliability for colour correction in all possible scenarios. A more sophisticated mechanism may be incorporated in colour enhancement pipeline. A useful cue would be the use of a number of other parameters like camera calibrations etc. to compute the threshold for the correction and enhancement of colours in the video frames. In the same way, the analysis of the colour histograms may be combined with the machine learning techniques to develop sophisticated model for the initial assessment of the quality of video frames in order to apply accurate correction and enhancement.

Another important future contribution would be the collection of a benchmark dataset that can be effectively used to compare various techniques for appearance based object retrieval. In literature, all the authors have used the images and videos recorded by themselves to report the results. These datasets have not been made available online and therefore other scientists cannot use them.

As mentioned in section 8.2 that the proposed vehicles make and model recognition system needs attention in optimization of the memory resource utilization. A potential direction would be the development of an efficient mechanism for processing only the selective number of visual words instead of all visual words. The procedure may be similar to the Bag of Visual Words (BoVW) approach where the most discriminative visual words are selected for the representation of the object. Similarly, an effective investigation for the dimensionality reduction of Mid Level Representation (MLR) based visual words is also important for performance improvement in terms of resource utilization and computational time. Also, the system needs to be tested and tuned for the rear view of the vehicles. The classification stage may also be enhanced and improved in the future research. Currently, the classification strategy relies on CS measure which may be replaced with a more sophisticated and faster classification technique to achieve the real-time performance.

The experiments conducted revealed a few limitations in the text detection and recognition system and opens up a few directions for future enhancements. The extraction of candidate regions in text detection stage is solely depending upon fixed number of colour quantization levels for all the images. A mechanism can be developed that analyses the colour distribution of the image and dynamically decides about the required quantization levels for a particular image. One helpful cue would be the total number of dominant colour centres in the image. The accuracy can be improved further if the shape information is also incorporated in the process of colour quantization. It will reduce the number of non-text regions in the initial candidate extraction stage. Currently, the initial screening of the candidate regions is being performed using only the aspect ratio, height and width. A number of other cues such as moment, edges, convex hull etc. may enhance the accuracy at the initial screening of candidate regions.

The process of classification of character binary map and background binary map can be made more accurate by adding another processing stage where the shape information can be used to improve the performance of the character identification. As mentioned in section 8.2, the proposed character identification technique fails in cases where the characters are joined either because of font style or because of variable lighting. A big room for improvement exists to improve the proposed method for such cases.

The character recognition stage has achieved exceptional performance. As a future work, more features may be tested to find out which feature representation is more suitable for text representation. In word recognition, a more reliable model may be tested to check the accuracy in conjunction with the proposed colour based character identification. In the same way, a more reliable criterion for the selection of text binary map and the background binary map in the word images is required. An important cue would be the use of shape information in both binary maps and the selection of the binary map as text binary map based on the patterns of the shape present in it.

References

[1] Abdel-Hakim, A. E., and Farag, A. A. Csift: A sift descriptor with color invariant characteristics. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on* (2006), vol. 2, pp. 1978–1983.

[2] Agarwal, A., and Triggs, B. Hyperfeatures - multilevel local coding for visual recognition. In *Proceedings of the 9th European Conference on Computer Vision - Volume Part I* (Berlin, Heidelberg, 2006), ECCV'06, Springer-Verlag, pp. 30–43.

[3] Akutsu, A., Tonomura, Y., Hashimoto, H., and Ohba, Y. Video indexing using motion vectors. vol. 1818, pp. 1522–1530.

[4] Bamberger, R., and Smith, M. A filter bank for the directional decomposition of images: theory and design. *Signal Processing, IEEE Transactions on 40*, 4 (Apr 1992), 882–893.

[5] Baudry, S., Chupeau, B., and Lefebvre, F. A framework for video forensics based on local and temporal fingerprints. In *Image Processing (ICIP), 2009 16th IEEE International Conference on* (Nov 2009), pp. 2889–2892.

[6] Bay, H., Tuytelaars, T., and Van Gool, L. Surf: Speeded up robust features. In *ECCV 2006*, A. Leonardis, H. Bischof, and A. Pinz, Eds., vol. 3951 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2006, pp. 404–417.

[7] Belongie, S., Malik, J., and Puzicha, J. Shape matching and object recognition using shape contexts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 24*, 4 (Apr 2002), 509–522.

[8] Betke, M., Haritaoglu, E., and Davis, L. Highway scene analysis in hard real-time. In *Intelligent Transportation System, 1997. ITSC '97., IEEE Conference on* (Nov 1997), pp. 812–817.

[9] Betke, M., and Nguyen, H. Highway scene analysis from a moving vehicle under reduced visibility conditions, 1998.

[10] Beymer, D., McLauchlan, P., Coifman, B., and Malik, J. A real-time computer vision system for measuring traffic parameters. In *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on* (Jun 1997), pp. 495–501.

[11] Bianco, S., Ciocca, G., Cusano, C., and Schettini, R. Improving color constancy using indoor outdoor image classification. *Image Processing, IEEE Transactions on 17*, 12 (Dec 2008), 2381–2392.

[12] Bissacco, A., Cummins, M., Netzer, Y., and Neven, H. Photoocr: Reading text in uncontrolled conditions. In *The IEEE International Conference on Computer Vision (ICCV)* (December 2013).

[13] Bradski, G. *Dr. Dobb's Journal of Software Tools* (2000).

[14] Bregler, C., and Malik, J. Learning appearance based models: Mixtures of second moment experts. In *NIPS* (1996), M. Mozer, M. I. Jordan, and T. Petsche, Eds., MIT Press, pp. 845–.

[15] Brown, L. M. Color retrieval for video surveillance. In *Advanced Video and Signal Based Surveillance, 2008. AVSS '08. IEEE Fifth International Conference on* (Sept 2008), pp. 283–290.

[16] Buchsbaum, G. A spatial processor model for object colour perception. *Journal of the Franklin Institute 310*, 1 (1980), 1 – 26.

[17] Burt, P. J., and Adelson, E. H. The laplacian pyramid as a compact image code. *IEEE Transactions on Communications 31* (1983), 532–540.

[18] Campos, T. d., Babu, B., and Verma, M. Character recognition in natural images. In *VISAPP* (2009).

[19] Carneiro, G., and Lowe, D. Sparse flexible models of local features. In *Proceedings of the 9th European Conference on Computer Vision - Volume Part III* (Berlin, Heidelberg, 2006), ECCV'06, Springer-Verlag, pp. 29–43.

[20] Chatfield, K., Lempitsky, V., Vedaldi, A., and Zisserman, A. The devil is in the details: an evaluation of recent feature encoding methods. In *Proceedings of the British Machine Vision Conference (BMVC)* (2011).

[21] Chen, H., Tsai, S. S., Schroth, G., Chen, D. M., Grzeszczuk, R., and Girod, B. Robust text detection in natural images with edge-enhanced maximally stable extremal regions. In *2011 IEEE International Conference on Image Processing* (Brussels, Sept. 2011).

[22] Chen, L., Hsieh, J. W., Yan, Y., and Wong, B. Y. Real-time vehicle make and model recognition from roads. In *2013 Conference on Information Technology and Applications in Outlying Islands.* (2013), pp. 128 – 132.

[23] Chen, T., Chen, Y., and Chien, S. Fast image segmentation based on k-means clustering with histograms in hsv color space. In *Multimedia Signal Processing, 2008 IEEE 10th Workshop on* (Oct 2008), pp. 322–325.

[24] Chen, X., and Yuille, A. L. Detecting and reading text in natural scenes. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Washington, DC, USA, 2004), CVPR'04, IEEE Computer Society, pp. 366–373.

[25] Chen, Y., and Chen, C. Fast human detection using a novel boosted cascading structure with meta stages. *IEEE Transactions on Image Processing 17*, 8 (2008), 1452–1464.

[26] Cieplinski, L. Mpeg-7 color descriptors and their applications. In *Computer Analysis of Images and Patterns*, W. Skarbek, Ed., vol. 2124 of *Lecture Notes in Computer Science.* Springer Berlin Heidelberg, 2001, pp. 11–20.

[27] Clady, X., Negri, P., Milgram, M., and Poulenard, R. Multi-class vehicle type recognition system. In *ANNPR* (2008), pp. 228–239.

[28] Courtney, J. D. Automatic video indexing via object motion analysis. *Pattern Recognition 30*, 4 (1997), 607 – 625.

[29] Csurka, G., Dance, C. R., Fan, L., Willamowski, J., and Bray, C. Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV* (2004), pp. 1–22.

[30] Cucchiara, R., Grana, C., Piccardi, M., and Prati, A. Detecting moving objects, ghosts and shadows in video streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence 25* (2003), 1337–1342.

[31] Cucchiara, R., Grana, C., Prati, A., and Vezzani, R. Probabilistic posture classification for human-behavior analysis. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on 35*, 1 (Jan 2005), 42–54.

[32] Dalal, N., and Triggs, B. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on* (June 2005), vol. 1, pp. 886–893 vol. 1.

[33] Daugman, J. G. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *J. Opt. Soc. Am. A 2*, 7 (Jul 1985), 1160–1169.

[34] Daya, B., Akoum, A. H., and Chauvet, P. Neural network approach for the identification system of the type of vehicle. In *Computational Intelligence and Communication Networks (CICN), 2010 International Conference on* (Nov 2010), pp. 162–166.

[35] Deb, K., Chae, H., and Jo, K. Vehicle license plate detection method based on sliding concentric windows and histogram. *JCP 4*, 8 (2009), 771–777.

[36] Do, M. N. Directional multiresolution image representations. Tech. rep., 2001.

[37] Do, M. N., and Vetterli, M. Framing pyramids. *IEEE Transactions on Signal Processing 51*, 9 (2003), 2329–2342.

[38] Do, M. N., and Vetterli, M. The contourlet transform: an efficient directional multiresolution image representation. *Image Processing, IEEE Transactions on 14*, 12 (Dec 2005), 2091–2106.

[39] Elgammal, A. M., Harwood, D., and Davis, L. S. Non-parametric model for background subtraction. In *Proceedings of the 6th European Conference on Computer Vision-Part II* (London, UK, UK, 2000), ECCV '00, Springer-Verlag, pp. 751–767.

[40] Epshtein, B., Ofek, E., and Wexler, Y. Detecting text in natural scenes with stroke width transform. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on* (June 2010), pp. 2963–2970.

[41] Ezaki, N., Bulacu, M., and Schomaker, L. Text detection from natural scene images: towards a system for visually impaired persons. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on* (Aug 2004), vol. 2, pp. 683–686 Vol.2.

[42] Feild, J. L., and Learned-Miller, E. G. Improving open-vocabulary scene text recognition. In *Proceedings of the 2013 12th International Conference on Document Analysis and Recognition* (Washington, DC, USA, 2013), ICDAR '13, IEEE Computer Society, pp. 604–608.

[43] Fisher, R. B. Pets04 surveillance ground truth data set. In *Sixth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS04), pp 1-5* (May 2004.).

[44] Forssen, P. E. Maximally stable colour regions for recognition and matching. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on* (June 2007), pp. 1–8.

[45] Fraz, M., Edirisinghe, E., and Sarfraz, M. Mid-level-representation based lexicon for vehicle make and model recognition. In *Pattern Recognition (ICPR), 2014 22nd International Conference on* (Aug 2014), pp. 393–398.

[46] Fraz, M., Sarfraz, M. S., and Edirisinghe, E. A. Exploiting color information for better scene text recognition. In *Proceedings of the British Machine Vision Conference* (2014), BMVA Press.

[47] Fraz, M., Sarfraz, M. S., and Edirisinghe, E. A. Exploiting colour information for better scene text detection and recognition. *International Journal on Document Analysis and Recognition (IJDAR) 18*, 2 (2015), 153–167.

[48] Fraz, M., Zafar, I., and Edirisinghe, E. A. Object colour extraction for cctv video annotation. In *Proceedings of the International Conference on Computer Vision Theory and Applications* (2013), pp. 455–459.

[49] Fraz, M., Zafar, I., Tzanidou, G., Edirisinghe, E. A., and Sarfraz, M. S. Human object annotation for surveillance video forensics. *Journal of Electronic Imaging 22*, 4 (2013), 041115–041115.

[50] Gonzalez, A., Bergasa, L., Yebes, J., and Bronte, S. Text location in complex images. In *Pattern Recognition (ICPR), 2012 21st International Conference on* (Nov 2012), pp. 617–620.

[51] Goyette, N., Jodoin, P., Porikli, F., Konrad, J., and Ishwar, P. Changedetection.net: A new change detection benchmark dataset. In *CVPR Workshops* (2012), IEEE, pp. 1–8.

[52] Hamdoun, O., Moutarde, F., Stanciulescu, B., and Steux, B. Person re-identification in multi-camera system by signature based on interest

point descriptors collected on short video sequences. In *Distributed Smart Cameras, 2008. ICDSC 2008. Second ACM/IEEE International Conference on* (Sept 2008), pp. 1–6.

[53] Hansen, D. M., Mortensen, B. K., Duizer, P. T., Andersen, J. R., and Moeslund, T. Automatic annotation of humans in surveillance video. In *Computer and Robot Vision, 2007. CRV '07. Fourth Canadian Conference on* (May 2007), pp. 473–480.

[54] Haselhoff, A., and Kummert, A. A vehicle detection system based on haar and triangle features. In *Intelligent Vehicles Symposium, 2009 IEEE* (June 2009), pp. 261–266.

[55] He, K., Sun, J., and Tang, X. Guided image filtering. In *Proceedings of the 11th European Conference on Computer Vision: Part I* (Berlin, Heidelberg, 2010), ECCV'10, Springer-Verlag, pp. 1–14.

[56] Heckbert, P. Color image quantization for frame buffer display. *SIGGRAPH Comput. Graph. 16*, 3 (July 1982), 297–307.

[57] Heikkila, J., and Silven, O. A real-time system for monitoring of cyclists and pedestrians. In *Visual Surveillance, 1999. Second IEEE Workshop on, (VS'99)* (Jul 1999), pp. 74–81.

[58] Hofmann, T. Probabilistic latent semantic indexing. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 1999), SIGIR '99, ACM, pp. 50–57.

[59] Hsieh, J., Chen, L., and Chen, D. Symmetrical surf and its applications to vehicle detection and vehicle make and model recognition. *Intelligent Transportation Systems, IEEE Transactions on 15*, 1 (Feb 2014), 6–20.

[60] Huang, T., Yang, G., and Tang, G. A fast two-dimensional median filtering algorithm. *Acoustics, Speech and Signal Processing, IEEE Transactions on 27*, 1 (Feb. 1979), 13–18.

[61] Huang, W., Lin, Z., Yang, J., and Wang, J. Text localization in natural images using stroke feature transform and text covariance descriptors. *Computer Vision, IEEE International Conference on 0* (2013), 1241–1248.

[62] Ilea, D. E., and Whelan, P. F. Color image segmentation using a spatial k-means clustering algorithm.

[63] J., M. I. Hierarchical mixtures of experts and the em algorithm. *Neural Computation 6* (1994), 181–214.

[64] Jaakkola, T., and Haussler, D. Exploiting generative models in discriminative classifiers. In *In Advances in Neural Information Processing Systems 11* (1998), MIT Press, pp. 487–493.

[65] Jain, A. K., and Zhong, Y. Page segmentation using texture analysis. *Pattern Recognition 29*, 5 (1996), 743 – 770.

[66] Javed, O., Shafique, K., and Shah, M. A hierarchical approach to robust background subtraction using color and gradient information. In *Motion and Video Computing, 2002. Proceedings. Workshop on* (Dec 2002), pp. 22–27.

[67] Jegou, H., Douze, M., Schmid, C., and Perez, P. Aggregating local descriptors into a compact image representation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on* (June 2010), pp. 3304–3311.

[68] Joachims, T. Text categorization with support vector machines: Learning with many relevant features. In *Machine Learning: ECML-98*, C. NÃľdellec and C. Rouveirol, Eds., vol. 1398 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 1998, pp. 137–142.

[69] Kaewtrakulpong, P., and Bowden, R. An improved adaptive background mixture model for realtime tracking with shadow detection. In *2nd European Workshop on Advanced Video Based Surveillance Systems* (2001).

[70] Kazemi, F., Samadi, S., Poorreza, H., and Akbarzadeh-T, M.-R. Vehicle recognition based on fourier, wavelet and curvelet transforms - a comparative study. In *Information Technology, 2007. ITNG '07. Fourth International Conference on* (April 2007), pp. 939–940.

[71] Kazemi, F., Samadi, S., Poorreza, H., and Akbarzadeh-T, M.-R. Vehicle recognition using curvelet transform and svm. In *Information Technology, 2007. ITNG '07. Fourth International Conference on* (April 2007), pp. 516–521.

[72] Kim, D., Kwak, J., Ko, B., and Nam, J. Human detection using wavelet-based cs-lbp and a cascade of random forests. In *Multimedia and Expo (ICME), 2012 IEEE International Conference on* (July 2012), pp. 362–367.

[73] Kovesi, P. Phase congruency: A low-level image invariant. *Psychological Research 64*, 2 (2000), 136–148.

[74] Kuo, J. Y., Lai, T. Y., Huang, F., and Liu, K. The color recognition of objects of survey and implementation on real-time video surveillance. In *SMC* (2010), IEEE, pp. 3741–3748.

[75] Lazebnik, S., Schmid, C., and Ponce, J. A sparse texture representation using affine-invariant regions. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on* (June 2003), vol. 2, pp. II–319–II–324 vol.2.

[76] Lazebnik, S., Schmid, C., and Ponce, J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on* (2006), vol. 2, pp. 2169–2178.

[77] Lee, C., Bharadwaj, A., Di, W., Jagadeesh, V., and Piramuthu, R. Region based descriminative pooling for scene text recognition. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on* (2014).

[78] Lee, C., Jung, K., and Kim, H. J. Automatic text detection and removal in video sequences. *Pattern Recognition Letters 24*, 15 (2003), 2607 – 2623.

[79] Lee, S., Gwak, J., and Jeon, M. Vehicle model recognition in video. *Signal Processing, Image Processing and Pattern Recognition 6*, 2 (April 2013).

[80] Leung, T., and Malik, J. Recognizing surfaces using three-dimensional textons. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on* (1999), vol. 2, pp. 1010–1017 vol.2.

[81] Li, X., Zhang, G., Fang, J., Wu, J., and Cui, Z. Vehicle color recognition using vector matching of template. In *Electronic Commerce and Security (ISECS), 2010 Third International Symposium on* (July 2010), pp. 189–193.

[82] Lim, Y., Choi, S., and Lee, S. Text extraction in mpeg compressed video for content-based indexing. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on* (2000), vol. 4, pp. 409–412 vol.4.

[83] Lipson, P., Grimson, E., and Sinha, P. Configuration based scene classification and image indexing. In *Proceedings of the 1997 Conference on*

Computer Vision and Pattern Recognition (CVPR '97) (Washington, DC, USA, 1997), CVPR '97, IEEE Computer Society, pp. 1007–.

[84] Lipson, P. R. *Context and Configuration-based Scene Classification.* PhD thesis, 1996. AAI0597725.

[85] Lipton, A., Clark, J., Brewer, P., Venetianer, P., and Chosak, A. Objectvideo forensics: activity-based video indexing and retrieval for physical security applications. In *Intelligent Distributed Surveilliance Systems, IEE* (Feb 2004), pp. 56–60.

[86] Lloyd, S. Least squares quantization in pcm. *IEEE Trans. Inf. Theor. 28*, 2 (Sept. 2006), 129–137.

[87] Lo, B. P. L., and Velastin, S. A. Automatic congestion detection system for underground platforms. In *Intelligent Multimedia, Video and Speech Processing, 2001. Proceedings of 2001 International Symposium on* (2001), pp. 158–161.

[88] Lowe, D. G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision 60*, 2 (Nov. 2004), 91–110.

[89] Lucas, S. M. Text locating competition results. In *ICDAR* (2005), IEEE Computer Society, pp. 80–85.

[90] Mao, W., Chung, F., Lam, K. K. M., and Siu, W. Hybrid chinese/english text detection in images and video frames. In *ICPR* (2002), pp. 1015–1018.

[91] Milyaev, S., Barinova, O., Novikova, T., Lempitsky, V., and Kohli, P. Image binarization for end-to-end text understanding in natural images. In *ICDAR* (2013), pp. 128 – 132.

[92] Mishra, A., Alahari, K., and Jawahar, C. V. An mrf model for binarization of natural scene text. In *ICDAR* (2011), pp. 11–16.

[93] Mishra, A., Alahari, K., and Jawahar, C. V. Scene text recognition using higher order language priors. In *Proceedings of the British Machine Vision Conference* (2012), BMVA Press, pp. 127.1–127.11.

[94] Mishra, A., Alahari, K., and Jawahar, C. V. Top-down and bottom-up cues for scene text recognition. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* (June 2012), pp. 2687–2694.

[95] Morrone, M., and Owens, R. Feature detection from local energy. *Pattern Recognition Letters 6*, 5 (1987), 303 – 313.

[96] Mosleh, A., Bouguila, N., and Hamza, A. B. Image text detection using a bandlet-based edge detector and stroke width transform. In *Proceedings of the British Machine Vision Conference* (2012), BMVA Press, pp. 63.1–63.12.

[97] Munroe, D. T., and Madden, M. G. Multi-Class and Single-Class Classification Approaches to Vehicle Model Recognition from Images. In *AICS* (2005).

[98] Navarro, G. A guided tour to approximate string matching. *ACM Comput. Surv. 33*, 1 (Mar. 2001), 31–88.

[99] Neumann, L., and Matas, J. A method for text localization and recognition in real-world images. In *Proceedings of the 10th Asian Conference on Computer Vision - Volume Part III* (Berlin, Heidelberg, 2011), ACCV'10, Springer-Verlag, pp. 770–783.

[100] Neumann, L., and Matas, J. Text localization in real-world images using efficiently pruned exhaustive search. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on* (Sept 2011), pp. 687–691.

[101] Neumann, L., and Matas, J. Real-time scene text localization and recognition. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* (June 2012), pp. 3538–3545.

[102] Neumann, L., and Matas, J. Scene text localization and recognition with oriented stroke detection. In *Computer Vision (ICCV), 2013 IEEE International Conference on* (Dec 2013), pp. 97–104.

[103] Novikova, T., Barinova, O., Kohli, P., and Lempitsky, V. Large-lexicon attribute-consistent text recognition in natural images. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part VI* (Berlin, Heidelberg, 2012), ECCV'12, Springer-Verlag, pp. 752–765.

[104] Nowak, E., Jurie, F., and Triggs, B. Sampling strategies for bag-of-features image classification. In *Computer Vision âĂŞ ECCV 2006*, A. Leonardis, H. Bischof, and A. Pinz, Eds., vol. 3954 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2006, pp. 490–503.

[105] Ojala, T., Pietikainen, M., and Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns.

Pattern Analysis and Machine Intelligence, IEEE Transactions on 24, 7 (Jul 2002), 971–987.

[106] Ojala, T., Pietikaineninen, M., and Harwood, D. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition 29*, 1 (1996), 51 – 59.

[107] Osuna, E., Freund, R., and Girosi, F. Training support vector machines: an application to face detection. In *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on* (Jun 1997), pp. 130–136.

[108] Pan, H., Li, X., Jin, L., and Xia, S. Contourlet-based feature extraction for object recognition. vol. 7495, pp. 749522–749522–8.

[109] Papageorgiou, C., Oren, M., and Poggio, T. A general framework for object detection. In *Computer Vision, 1998. Sixth International Conference on* (Jan 1998), pp. 555–562.

[110] Papageorgiou, C., and Poggio, T. A trainable system for object detection. *Int. J. Comput. Vision 38*, 1 (June 2000), 15–33.

[111] Payne, A. M., Bhaskar, H., and Mihaylova, L. Multi-resolution learning vector quantisation based automatic colour clustering. In *Information Fusion, 2008 11th International Conference on* (June 2008), pp. 1–6.

[112] Pearce, G., and Pears, N. Automatic make and model recognition from frontal images of cars. In *Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on* (Aug 2011), pp. 373–378.

[113] Perona, P., and Malik, J. Scale-space and edge detection using anisotropic diffusion. *IEEE Trans. Pattern Anal. Mach. Intell. 12*, 7 (July 1990), 629–639.

[114] Perronnin, F., Sanchez, J., and Mensink, T. Improving the fisher kernel for large-scale image classification. In *Proceedings of the 11th European Conference on Computer Vision: Part IV* (Berlin, Heidelberg, 2010), ECCV'10, Springer-Verlag, pp. 143–156.

[115] Petrovic, V., and Cootes, T. F. Analysis of features for rigid structure vehicle type recognition. In *In British Machine Vision Conference* (2004), pp. 587–596.

[116] Phan, T. Q., Shivakumara, P., Tian, S., and Tan, C. L. Recognizing text with perspective distortion in natural scenes. In *ICCV* (2013), pp. 569–576.

[117] Prokaj, J., and Medioni, G. 3-d model based vehicle recognition. In *Applications of Computer Vision (WACV), 2009 Workshop on* (Dec 2009), pp. 1–7.

[118] Psyllos, A., Anagnostopoulos, C., and Kayafas, E. Vehicle model recognition from frontal view image measurements. *Computer Standards & Interfaces 33*, 2 (2011), 142 – 151.

[119] Rahati, S., Moravejian, R., Mohamad, E., and Mohamad, F. Vehicle recognition using contourlet transform and svm. In *Information Technology: New Generations, 2008. ITNG 2008. Fifth International Conference on* (April 2008), pp. 894–898.

[120] Ramnath, K., Sinha, S. N., Szeliski, R., and Hsiao, E. Car make and model recognition using 3d curve alignment. In *IEEE Winter Conference on Applications of Computer Vision (WACV 2014)* (March 2014), IEEE Computer Society.

[121] Saber, E., Tekalp, A. M., Eschbach, R., and Knox, K. Automatic image annotation using adaptive color classification. *GRAPHICAL MODELS AND IMAGE PROCESSING 58*, 2 (1999), 115–126.

[122] Sanchez, J., and Perronnin, F. High-dimensional signature compression for large-scale image classification. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on* (June 2011), pp. 1665–1672.

[123] Saravi, S., and Edirisinghe, E. Vehicle make and model recognition in cctv footage. In *Digital Signal Processing (DSP), 2013 18th International Conference on* (July 2013), pp. 1–6.

[124] Sarfraz, M., Shahzad, A., Elahi, M. A., Fraz, M., Zafar, I., and Edirisinghe, E. Real-time automatic license plate recognition for cctv forensic applications. *Journal of Real-Time Image Processing 8*, 3 (2013), 285–295.

[125] Sarfraz, M. S., and Hellwich, O. Head pose estimation in face recognition across pose scenarios. In *VISAPP* (2009).

[126] Sarfraz, M. S., and Khan, M. H. A probabilistic framework for patch based vehicle type recognition. In *VISAPP* (2011), L. Mestetskiy and J. Braz, Eds., SciTePress, pp. 358–363.

[127] Sarfraz, M. S., Saeed, A., Khan, M. H., and Riaz, Z. Bayesian prior models for vehicle make and model recognition. In *Proceedings of the 7th International Conference on Frontiers of Information Technology* (New York, NY, USA, 2009), FIT '09, ACM, pp. 35:1–35:6.

[128] Schauerte, B., and Fink, G. A. Web-based learning of naturalized color models for human-machine interaction. In *Proceedings of the 12th International Conference on Digital Image Computing: Techniques and Applications (DICTA)* (Sydney, Australia, December 1-3 2010), IEEE.

[129] Schauerte, B., and Stiefelhagen, R. Learning robust color name models from web images. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR)* (Tsukuba, Japan, November 11-15 2012), IEEE.

[130] Schettini, R., Ciocca, G., and Zuffi, C. A survey of methods for colour image indexing and retrieval in image databases. In *IN COLOR IMAGING SCIENCE: EXPLOITING DIGITAL* (2001), Media, John Wiley, pp. 9–1.

[131] Schindler, K., and Wang, H. Smooth foreground-background segmentation for video processing. In *Computer Vision âĂŞ ACCV 2006*, P. Narayanan, S. Nayar, and H.-Y. Shum, Eds., vol. 3852 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2006, pp. 581–590.

[132] Schwartz, W., and Davis, L. Learning discriminative appearance-based models using partial least squares. In *Computer Graphics and Image Processing (SIBGRAPI), 2009 XXII Brazilian Symposium on* (Oct 2009), pp. 322–329.

[133] Scott, D. W. Multivariate density estimation and visualization. In *Handbook of Computational Statistics*, J. E. Gentle, W. Haerdle, and Y. Mori, Eds. Springer-Verlag, Berlin, 2004, pp. 517–538. Chapter III.4.

[134] Shahab, A., Shafait, F., and Dengel, A. Icdar 2011 robust reading competition challenge 2: Reading text in scene images. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on* (Sept 2011), pp. 1491–1496.

[135] Shao, H., Wu, Y., Cui, W., and Zhang, J. Image retrieval based on mpeg-7 dominant color descriptor. In *Young Computer Scientists, 2008. ICYCS 2008. The 9th International Conference for* (Nov 2008), pp. 753–757.

[136] Shengcai, L., and Jain, A. K. Partial face recognition: An alignment free approach. In *Biometrics (IJCB), 2011 International Joint Conference on* (Oct 2011), pp. 1–8.

[137] Sheshadri, K., and Divvala, S. K. Exemplar driven character recognition in the wild. In *British Machine Vision Conference (BMVC) 2012* (September 2012).

[138] Shi, C., Wang, C., Xiao, B., Z., Y., and Gao, S. Scene text detection using graph model built upon maximally stable extremal regions. *Pattern Recognition Letters 34*, 2 (2013), 107 – 116.

[139] Shi, C., Wang, C., Xiao, B., Zhang, Y., Gao, S., and Zhang, Z. Scene text recognition using part-based tree-structured character detection. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on* (June 2013), pp. 2961–2968.

[140] Shivakumara, P., Phan, T. Q., and Tan, C. A laplacian approach to multi-oriented text detection in video. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 33*, 2 (Feb 2011), 412–419.

[141] Silverman, B. W. *Density Estimation for Statistics and Data Analysis.* Chapman & Hall, London, 1986.

[142] Simonyan, K., Parkhi, O. M., Vedaldi, A., and Zisserman, A. Fisher Vector Faces in the Wild. In *British Machine Vision Conference* (2013).

[143] Sin, B., Kim, S., and Cho, B. Locating characters in scene images using frequency features. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on* (2002), vol. 3, pp. 489–492 vol.3.

[144] Smith, D. L., Field, J., and Learned-Miller, E. Enforcing similarity constraints with integer programming for better scene text recognition. *2013 IEEE Conference on Computer Vision and Pattern Recognition 0* (2011), 73–80.

[145] Smith, J. R., and Chang, S.-F. Single color extraction and image query. In *Proc. IEEE International Conference on Image Processing* (1995).

[146] Son, J., Park, S., and Kim, K. A convolution kernel method for color recognition. In *Advanced Language Processing and Web Information Technology, 2007. ALPIT 2007. Sixth International Conference on* (Aug 2007), pp. 242–247.

[147] Sosa, L. P., Lucas, S. M., Panaretos, A., Sosa, L., Tang, A., Wong, S., and Young, R. Icdar 2003 robust reading competitions. In *In Proceedings of the Seventh International Conference on Document Analysis and Recognition* (2003), IEEE Press, pp. 682–687.

[148] Stauffer, C., and Grimson, W. E. L. Adaptive background mixture models for real-time tracking. In *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.* (Los Alamitos, CA, USA, Aug. 1999), vol. 2, IEEE, pp. 246–252 Vol. 2.

[149] Stauffer, C., and Grimson, W. E. L. Learning patterns of activity using real-time tracking. *IEEE Trans. Pattern Anal. Mach. Intell. 22*, 8 (Aug. 2000), 747–757.

[150] Swain, M. J., and Ballard, D. H. Color indexing. *Int. J. Comput. Vision 7*, 1 (Nov. 1991), 11–32.

[151] Tkalcic, M., and Tasic, J. Colour spaces: perceptual, historical and applicational background. In *EUROCON 2003. Computer as a Tool. The IEEE Region 8* (Sept 2003), vol. 1, pp. 304–308 vol.1.

[152] Tomasi, C., and Manduchi, R. Bilateral filtering for gray and color images. In *Proceedings of the Sixth International Conference on Computer Vision* (Washington, DC, USA, 1998), ICCV '98, IEEE Computer Society, pp. 839–.

[153] Turk, M., and Pentland, A. Eigenfaces for recognition. *J. Cognitive Neuroscience 3*, 1 (Jan. 1991), 71–86.

[154] Tuzel, O., and Meer, P. Region covariance: A fast descriptor for detection and classification. In *In Proc. 9th European Conf. on Computer Vision* (2006), pp. 589–600.

[155] Van de Weijer, J., Gevers, T., and Gijsenij, A. Edge-based color constancy. *IEEE Transactions on Image Processing 16*, 9 (2007), 2207–2214.

[156] Van de Weijer, J., Schmid, C., and Verbeek, J. Learning color names from real-world images. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on* (June 2007), pp. 1–8.

[157] Vapnik, V. N. The nature of statistical learning theory, 1995.

[158] Vapnik, V. N. *Statistical Learning Theory.* Wiley, 1998.

[159] Varma, M., and Zisserman, A. A statistical approach to texture classification from single images. *Int. J. Comput. Vision 62*, 1-2 (Apr. 2005), 61–81.

[160] Vedaldi, A., and Fulkerson, B. Vlfeat: An open and portable library of computer vision algorithms. In *Proceedings of the International Conference on Multimedia* (New York, NY, USA, 2010), MM '10, ACM, pp. 1469–1472.

[161] Vetterli, M. Multi-dimensional sub-band coding: Some theory and algorithms. *Signal Processing 6*, 2 (1984), 97 – 112.

[162] Viola, P., and Jones, M. J. Robust real-time face detection. *Int. J. Comput. Vision 57*, 2 (May 2004), 137–154.

[163] Viola, P., Jones, M. J., and Snow, D. Detecting pedestrians using patterns of motion and appearance. *Int. J. Comput. Vision 63*, 2 (July 2005), 153–161.

[164] Wakahara, T., and Kita, K. Binarization of color character strings in scene images using k-means clustering and support vector machines. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on* (Sept 2011), pp. 274–278.

[165] Wang, H., and Suter, D. A re-evaluation of mixture of gaussian background modeling [video signal processing applications]. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on* (2005), vol. 2, pp. ii/1017–ii/1020 Vol. 2.

[166] Wang, H., and Suter, D. Tracking and segmenting people with occlusions by a sample consensus based method. In *ICIP (2)* (2005), IEEE, pp. 410–413.

[167] Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., and Gong, Y. Locality-constrained linear coding for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on* (June 2010), pp. 3360–3367.

[168] Wang, K., Babenko, B., and Belongie, S. End-to-end scene text recognition. In *IEEE International Conference on Computer Vision (ICCV)* (Barcelona, Spain, 2011).

[169] Wang, K., and Belongie, S. Word spotting in the wild. In *European Conference on Computer Vision (ECCV)* (Heraklion, Crete, Sept. 2010).

[170] Wang, X., Doretto, G., Sebastian, T., Rittscher, J., and Tu, P. Shape and appearance context modeling. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on* (Oct 2007), pp. 1–8.

[171] Wang, X., Han, T. X., and Yan, S. An hog-lbp human detector with partial occlusion handling. In *ICCV* (2009), IEEE, pp. 32–39.

[172] Weinman, J. J., Butler, Z., Knoll, D., and Feild, J. Toward integrated scene text reading. *IEEE Transactions on Pattern Analysis and Machine Intelligence 36*, 2 (2014), 375–387.

[173] Winn, J., Criminisi, A., and Minka, T. Object categorization by learned universal visual dictionary. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on* (Oct 2005), vol. 2, pp. 1800–1807 Vol. 2.

[174] Wolf, C., and Jolion, J. Object count/area graphs for the evaluation of object detection and segmentation algorithms. *IJDAR 8*, 4 (2006), 280–296.

[175] Wren, C. R., Azarbayejani, A., Darrell, T., and Pentland, A. P. Pfinder: real-time tracking of the human body. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 19*, 7 (Jul 1997), 780–785.

[176] Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S., and Yi, M. Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 31*, 2 (Feb 2009), 210–227.

[177] Wu, G., Rahimi, A., Chang, E., Goh, K., Tsai, T., Jain, A., and Wang, Y. Identifying color in motion in video sensors. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on* (June 2006), vol. 1, pp. 561–569.

[178] Wu, V., Manmatha, R., and Riseman, E. M. Finding text in images, 1997.

[179] Wu, V., Manmatha, R., and Riseman, E. M. Textfinder: An automatic system to detect and recognize text in images, 1997.

[180] Yao, C., Bai, X., Liu, W., Ma, Y., and Tu, Z. Detecting texts of arbitrary orientations in natural images. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* (June 2012), pp. 1083–1090.

[181] Yao, C., Bai, X., Shi, B., and Liu, W. Strokelets: A learned multi-scale representation for scene text recognition.

[182] Ye, Q., Huang, Q., Gao, W., and Zhao, D. Fast and robust text detection in images and video frames. *Image Vision Comput. 23*, 6 (June 2005), 565–576.

[183] Yi, C., and Tian, Y. Text string detection from natural scenes by structure-based partition and grouping. *Image Processing, IEEE Transactions on 20*, 9 (Sept 2011), 2594–2605.

[184] Yi, C., and Tian, Y. Text extraction from scene images by character appearance and structure modeling. *Comput. Vis. Image Underst. 117*, 2 (Feb. 2013), 182–194.

[185] Yi, C., Yang, X., and Tian, Y. Feature representations for scene text character recognition: A comparative study. In *Proceedings of the 2013 12th International Conference on Document Analysis and Recognition* (Washington, DC, USA, 2013), ICDAR '13, IEEE Computer Society, pp. 907–911.

[186] Zafar, I., Edirisinghe, E. A., and Acar, B. S. Localized contourlet features in vehicle make and model recognition. vol. 7251, pp. 725105–725105–9.

[187] Zafar, I., Fraz, M., and Edirisinghe, E. A. Human object articulation for cctv video forensics. vol. 8663, pp. 86630Z–86630Z–11.

[188] Zhang, Y., Chou, C. C., Yu, S. S., and Chen, T. Object color categorization in surveillance videos. In *18th IEEE International Conference on Image Processing, ICIP 2011, Brussels, Belgium, September 11-14, 2011* (2011), pp. 2913–2916.

[189] Zhang, Y., Yu, S. S., and Chen, T. Improving object color categorization with shapes. In *Proceedings of the International Conference on Image Processing, ICIP 2010, September 26-29, Hong Kong, China* (2010), pp. 1053–1056.

[190] Zheng, Y., Shen, C., Hartley, R. I., and Huang, X. Effective pedestrian detection using center-symmetric local binary/trinary patterns. *CoRR abs/1009.0892* (2010).

[191] Zheng, Y., Shen, C., and Huang, X. Pedestrian detection using center-symmetric local binary patterns. In *ICIP* (2010), IEEE, pp. 3497–3500.

[192] Zhong, Y., Zhang, H., and Jain, A. K. Automatic caption localization in compressed video. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 22*, 4 (Apr 2000), 385–392.

[193] Zhou, Q., and Aggarwal, J. K. Tracking and classifying moving objects from videos, 2001.

[194] Zhou, X., Yu, K., Zhaou, K., and Huang, T. S. Image classification using super-vector coding of local image descriptors. In *Computer Vision âĂŞ ECCV 2010*, K. Daniilidis, P. Maragos, and N. Paragios, Eds., vol. 6315 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2010, pp. 141–154.