Title	Image and video quality in the automotive environment
Author(s)	Winterlich, John Anthony
Publication Date	2016-05-12
Item record	http://hdl.handle.net/10379/5782

Image and Video Quality in the Automotive Environment

A thesis presented
by

John Anthony Winterlich

to
The College of Engineering & Informatics
in fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of

Electrical & Electronic Engineering

National University of Ireland, Galway
Galway, Ireland
May 2016
Supervisor: Dr. Edward Jones
Co-Supervisor: Dr. Martin Glavin
Professor in Discipline: Prof. Gerard Hurley

Abstract

This thesis is concerned with the development of methodologies for image and video quality assessment for the automotive environment, and the use of those methodologies for evaluating the effect of image and video degradations in automotive vision systems. Image quality metrics are important tools for optimizing system design parameters associated with image acquisition, compression and transmission. While optimizing systems for perceptual quality is already a common element of consumer electronics devices, in the automotive environment Advanced Driver Assistance Systems (ADAS) incorporating machine vision applications such as automated pedestrian detection are becoming a more widely-used feature of vehicular vision systems and the quality requirements of such systems present an additional challenge. As such, automotive image quality must also be tuned for optimal machine vision performance.

In this thesis, quality is considered from the perspective of both machine vision performance and perceptual quality. An evaluation of the effect of image degradations on pedestrian detection performance is first carried out. This study highlights the quality impact of different imaging system degradations, such as compression artifacts, on pedestrian detection performance. It is demonstrated that improvements in detection performance can be achieved by training detection algorithms on images with a wide variety of degradations. A full-reference objective image quality assessment algorithm based on Histograms of Oriented Gradients (HOGs) is also proposed that correlates closely with pedestrian detection performance on degraded video frames. A system for No-Reference distortion classification, suitable for real-time operation, is also proposed. The classification system, based on natural image statistics, is combined with a multi-classifier approach to pedestrian detection in order to increase pedestrian detection performance on degraded images. Furthermore, a new approach for predicting the subjective Quality of Experience (QoE) of fish-eye to rectilinear transformed images is proposed. Improved correlation with subjective human opinion is achieved by weighting local quality scores with saliency information. Finally, an automotive specific video quality database is presented consisting of 50 video sequences and associated human saliency data and mean opinion scores. The influence of packet loss on visual QoE for high bandwidth automotive networks is also considered. The results show that increasing the level of packet loss has almost no effect on visual attention, despite significant differences in the MOS scores with different levels of packet loss.

Table of Contents

I hereby declare that the work contained in this thesis has not been submitted by me in pursuance of any other degree.

Name:

Date:

List of Figures

List of Tables

Sponsor Acknowledgment

This research was supported by funding from the Irish Research Council (IRC) and Valeo Vision Systems under the IRC Enterprise Partnership Scheme.

Acknowledgments

Completing this doctorate has been a demanding experience, which I could not have completed without the support of my supervisors, colleagues, friends and family. Firstly to my parents, Séan and Helen, for always supporting and encouraging me in all of my endeavours, I cannot thank them enough. Many thanks also to my brothers Gary and Derek, and my sister Gráinne for their support.

I would like to sincerely thank my PhD supervisor Dr. Edward Jones, co-supervisor Dr. Martin Glavin, and Mr. Liam Kilmartin. Throughout my PhD they were always on-hand to to offer advice, direction and constructive criticism. Completing this thesis would not have been possible without their guidance and support.

Carrying out research with an industry focus was a particularly challenging experience. My industry partners Valeo Vision Systems offered a tremendous amount of technical support throughout my research. In particular I would like to thank Dr. Patrick Denny, Dr. Ciarán Hughes, and Dr. Vladimir Zlokolica for their support.

I have often heard it said that completing a PhD can be a lonely experience. My experience has been quite different. I have had difficult times, certainly - rejected papers and poor experimental results. I've also had extremely positive experiences - winning research awards and publishing featured articles. Throughout all of these experiences I have been extremely fortunate to share an office with a group of researchers that have shared my triumphs and adversities, kept me grounded when I was successful, and picked me back up when I failed. The camaraderie and good humour of all the researchers in the Electronic Engineering discipline has made my PhD an enjoyable and rewarding experience. In particular I would like to thank the current members of the Connaught Automotive Research Group: Shane Tuohy, Pat

Hurney, Damien Dooley, Martin Gallagher and Brian McGinley for their friendship and support.

Lastly, but most of all I would like to thank my wife Jinhee. For her love, kindness and support I am forever grateful. She is a constant source of inspiration to me. I would not have completed this PhD without her.

Glossary of Terms

ACR-HR: Absolute Category Rating with Hidden Reference

ADAS: Advanced Driver Assistance Systems

AP: Average Precision

AUC: Area Under Curve

AWGN: Additive White Gaussian Noise

BLIINDS: BLind Image Integrity Notator using DCT Statistics

BRISQUE: Blind/Referenceless Image Spatial QUality Evaluator

CCD: Charge Coupled Device

CCH: Chain Code Histogram

CD: Change Detection

CMOS: Complementary Metal-Oxide Semiconductor

CR: Compression Ratio

DCT: Discrete Cosine Transfer

DESIQUE: DErivative Statistics-based Image QUality Evaluator

DIIVINE: Distortion Identification-based Image Verity and INtegrity Evaluation

DMOS: Differential Mean Opinion Score

DNT: Divisive Normalisation Transfer

DPM: Deformable Parts Model

FM: Fixation Map

FOV: Field of View

FPPI: False Positives Per Image

FR: Full-Reference

FSIM: Feature SIMilarity

GBVS: Graph Based Visual Saliency

GGD: Generalised Gaussian Distribution

GM: Gradient Magnitude

GSM: Gaussian Scale Mixtures

HDR: High Dynamic Range

HMSE: HOG-Mean Squared Error

HOG: Histogram of Oriented Gradients

HVS: Human Visual System

ICF: Integral Channel Features

IFC: Information Fidelity Criterion

INRIA: Institut National de Recherche en Informatique et en Automatique

IP: Internet Protocol

IQA: Image Quality Assessment

ITU: International Telecommunications Union

IW-SSIM: Information-content Weighted SSIM

JP2: JPEG 2000

JPEG: Joint Photographic Experts Group

KL: Kullback-Leibler

LAMR: Log-Average Miss Rate

LBP: Local Binary Patterns

LIVE: Laboratory for Image and Vision Engineering

MLSIM: Multi-level SIMilarity

MOS: Mean Opinion Score

MSE: Mean Squared Error

MVG: Multi-Variate Gaussian

NIQE: Natural Image Quality Evaluator

NQM: Noise Quality Metric

NR: No-Reference

NS: Normalized Similarity

NSS: Natural Scene Statistics

PASCAL: Pattern Analysis, Statistical Modelling and Computational Learning

PC: Phase Congruency

PGH: Pairwise Geometrical Histogram

PLCC: Pearson's Linear Correlation Coefficient

PSNR: Peak Signal to Noise Ratio

QoE: Quality of Experience

RMSE: Root Mean Squared Error

ROC: Receiver Operating Characteristic

RR: Reduced-Reference

SI: Spatial Information

SIFT: Scale Invariant Feature Transform

SM: Saliency Map

SR-SIM: Spectral Residual SIMilarity

SSIM: Structural SIMilarity

SVM: Support Vector Machine

TI: Temporal Information

UESL: Upper Empirical Similarity Limit

VIF: Visual Image Fidelity

VQEG: Video Quality Experts Group

VRU: Vulnerable Road User

VSNR: Visual Signal to Noise Ratio

Chapter 1

Introduction

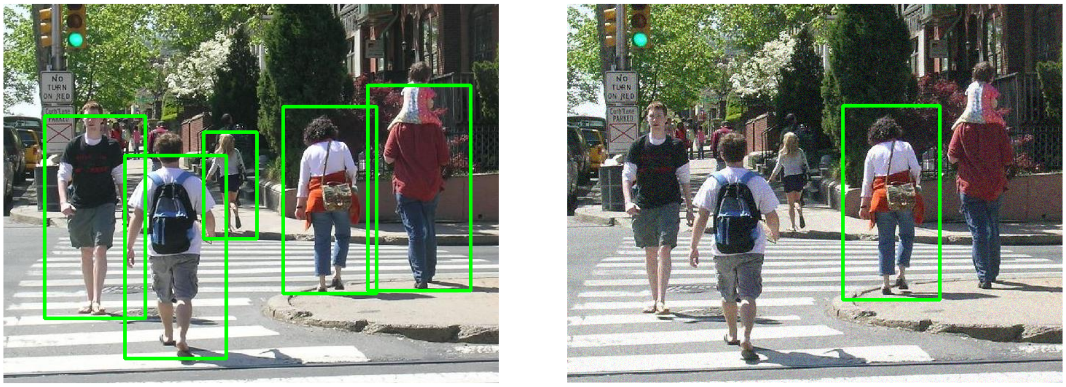
1.1 Motivation

Although road safety in Europe has improved considerably in the last 20 years [1] it remains a major societal issue, with an unacceptable number of fatalities occurring on our roads every year. For example, in 2011, more than 30,000 people died from road accidents in the European Union [2], equivalent to the population of a medium sized town. Moreover, for every death on European roads there are an estimated 4 permanently disabling injuries such as damage to the spinal cord or the brain. In total, well over a million people are injured on European roads every year, the economic cost of which has been estimated as 130 billion euros [1]. In 2010, the EU commission adopted an ambitious road safety program aimed at reducing European road deaths by 50% by the year 2020 [1]. Making vehicles safer is an important component of efforts to reduce road traffic injuries and many technologies are being applied to prevent crashes. For example, anti-skid electronic stability control is now increasingly required as a mandatory safety feature for new passenger cars and light duty vehicles. Measures intended to reduce the risk and severity of pedestrian impact are also becoming important in vehicle design, with computer vision systems increasingly being used for driver assistance [3]. Vision systems may simply provide additional visual information in situations where the driver's view may be obstructed, for example when reverse parking or encountering cross traffic situations with limited visibility. However in many cases Advanced Driver Assistance Systems (ADAS) incorporate sophisticated machine vision software that actively processes video data and works to mitigate collisions [4]. An important class of ADAS is the detection of Vulnerable Road Users (VRUs), who are defined in the European Commission's intelligent transport systems directive as: "non-motorised road users, such as pedestrians and cyclists as well as motor-cyclists and persons with disabilities or reduced mobility and orientation" [5]. In particular, the task of pedestrian detection from cameras mounted on a vehicle has become increasingly important to a number of automotive safety applications such as collision avoidance and autonomous driving [6, 7, 8, 9, 10].

Clearly, automotive vision systems are safety critical, and hence the quality of video displayed to the driver, or processed by a machine vision algorithm is of paramount importance, but despite the fact that research into ADAS has grown significantly in recent years, a surprisingly small amount of research has examined the impact of image degradations on driver assistance system performance. In the automotive environment this is a topic of considerable importance since imperfections in visual quality can occur in a variety of ways. Typical automotive vision systems use Complementary Metal-Oxide-Semiconductor (CMOS) sensors with ultra wide fields of view. Degradations in image quality can occur as a result of sensor noise at image capture, lossy compression, or transmission errors [11, 12, 13]. Evaluating the influence of such degradations on the quality of automotive systems is an important step towards developing robust driver assistance systems. At present, while there are methods in existence for subjective and objective image and video quality assessment that are widely applied in consumer applications, there are no accepted industry standards or tools specifically for the automotive environment. This thesis aims to address this gap by developing tools and techniques to quantify the quality of video used in automotive vision systems.

Typically, in order to quantify the perceptual quality of an image or video sequence, subjective tests must be carried out in which a sizeable number of human observers are shown a series of images or video sequences whose quality they are asked to rate on a particular scale [14]. The mean score of each image is termed the Mean Opinion Score (MOS) and is representative of the perceived quality of that image or video. Subjective tests can often provide reliable assessments of quality since they may be designed to accurately represent a specific application. Such large scale subjective tests have been carried out for generic video data and the results have been made publicly available to the research community, for example in [15, 16, 17]. Methodologies for subjective assessment of video quality are very well described by the International Telecommunications Union (ITU) for certain applications, such as television broadcasting [14] or multimedia applications [18]. However, the concept of video quality for automotive vision systems differs greatly from that of entertainment based consumer image quality, since in the automotive environment, the subjective satisfaction of the user depends upon achieving a particular task, such as event detection or object recognition [19, 20]. Unfortunately, there is no clear definition of quality in the context of automotive vision systems, even if one limits consideration to quality as perceived by a human observer, i.e. the driver. In [21], the authors make a distinction between the "naturalness" and "usefulness" of an automotive image. The "naturalness" of an automotive image is generally accepted as being closely related to the traditional notion of perceptual image quality. A natural image should be a faithful representation of the road ahead. For example, it should contain recognizable signal colours and be free from noise or compression artifacts. On the other hand, the same image with exaggerated local contrast and sharpness may be more "useful" if it allows the driver to see more detail, such as a pedestrian on a dark street. In an automotive context, the amount of information that can be extracted from a scene determines its "usefulness" [21]. For this reason, a typical automotive vision system will employ the use of fish-eye lenses with Fields Of View (FOV) of up to 190 degrees. These lenses are more useful as they enable drivers to see more objects approaching from the sides. The enhanced FOV is also convenient for driver assistance technologies such as side collision warning systems. Fish-eye lenses, though offering an enhanced FOV, introduce significant non-linear radial distortion to an image, which is manifested in straight lines being mapped to curves. Furthermore, the perspective of projection of a given scene in the fish-eye view differs greatly from the projection of the same scene in a rectilinear pin-hole camera, thus making it difficult for drivers to accurately judge distances to other vehicles and pedestrians. To mitigate this problem, automotive fish-eye images are often partially or wholly corrected for radial distortion. The use of polynomial models to correct radially distorted images to the more familiar rectilinear image form is well established [22]. Polynomial mapping can in turn introduce interpolation artifacts to the transformed image. Due to the fact that radial image resolution gradually decreases from the centre of the image to its peripheral areas, blurring in the peripheral areas of the images can also occur [23]. These distortions can significantly reduce the perceived visual quality of the transformed rectilinear image in cases where the focus of interest is in the area of reduced spatial resolution, however to what extent they reduce the "usefulness" of such images for object recognition or the completion of a particular task, such as reverse parking, is unclear. In order to accurately predict quality for recognition of objects of interest in the scene, it is necessary to consider the scene content of each image or video, paying particular attention to regions of high visual saliency. However, despite the fact that fish-eye lenses are becoming increasingly popular in applications such as video surveillance, robotics and automotive vision systems, little subjective or objective assessment of the perceptual quality of these images has been carried out.

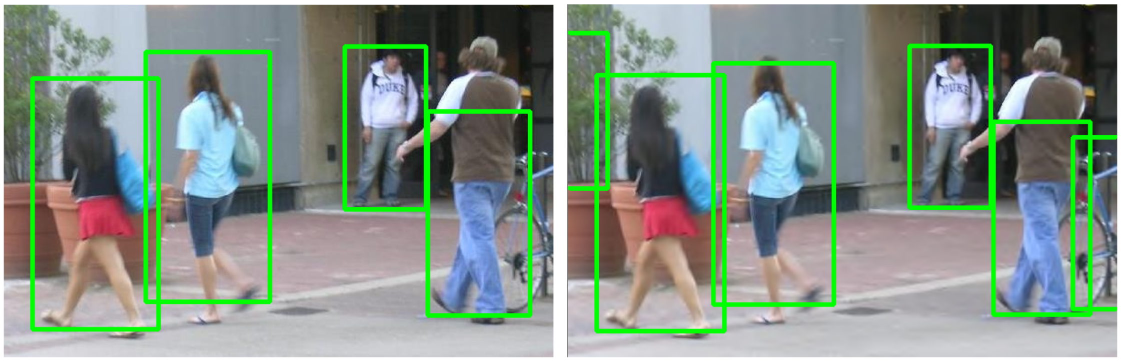
Although the distinction between naturalness and usefulness is important when considering quality as perceived by the driver, the situation becomes more complicated when considering machine vision. Increasingly, image processing algorithms are making use of automotive cameras for applications such as automatic pedestrian or vehicle detection. The question then arises as to whether an image that is useful to the human driver is equally useful to a machine vision algorithm. Unfortunately, this is often not the case. For example, depending on the content of an image, the addition



(a) (b)

Figure 1.1: Impact of noise on human vs. machine vision. Figure (a) is an original clean image. Green boxes indicate detection of pedestrians by a high performance machine vision algorithm. The presence of noise in (b) can go unnoticed to a human viewer due to masking effects, whereas noise can have a significant effect on the performance of the pedestrian detection algorithm.

of noise may go unnoticed to a human viewer due to masking effects, whereas even a modest amount of noise will typically impair the performance of image processing algorithms (as shown in Figure 1.1). Similarly, interpolation artifacts may occur at high frequencies which are difficult for a human viewer to discern (as illustrated in Figure 1.2), nevertheless the presence of such artifacts are sufficient to degrade the performance of a pedestrian detection algorithm. These examples raise important safety concerns for automotive vision engineers since a seemingly imperceptible change in image quality can significantly alter the performance of machine vision algorithms. While optimizing systems for perceptual quality is important, systems must also be tuned for optimal machine vision performance. In this thesis it is demonstrated that the human visual system (HVS) may not always perceive image distortions that adversely affect machine vision performance and as such, existing Image Quality Assessment (IQA) algorithms are not necessarily reliable predictors of machine vision performance on transmitted video sequences, prompting the need to measure quality in a way that is meaningful for machine vision algorithms, and stimulating the development of more appropriate algorithms for the automotive environment. This is the primary focus of this thesis.



(a) (b)

Figure 1.2: (a) Original clean image. “Ringing” caused by JPEG2000 (JP2) compression (b) can occur at high spatial frequencies which are difficult for the human visual system to discern, nonetheless the presence of this distortion is sufficient to degrade the performance of a pedestrian detection algorithm by causing an increase in false positives (indicated by spurious green boxes).

1.2 Contributions of this Thesis

1.2.1 Contributions

This thesis is concerned with the development of methodologies for image and video quality assessment appropriate to the automotive environment, and the use of those methodologies for evaluating the effect of image and video degradations in automotive vision systems. More specifically, the primary contributions of this thesis are as follows:

1. An evaluation of the effect of image degradations on pedestrian detection performance has been carried out. This study highlights the quality impact of different design decisions, such as compression levels, on pedestrian detection performance. It is demonstrated that improvements in detection performance can be achieved by training detection algorithms on images with a wide variety of degradations.

2. A Full-Reference (FR) IQA metric based on Histograms of Oriented Gradients (HOGs) is proposed. The metric accurately predicts the performance of pedestrian detection algorithms on degraded images.

3. A system for No-Reference (NR) distortion classification is proposed. The classification metric, based on natural image statistics, is combined with a multi-classifier approach to pedestrian detection in order to increase detection performance on degraded images. The proposed system enhances the pedestrian detection performance of existing methods and has the potential to be used

in real-time in-vehicle networks to improve pedestrian detection performance across a wide range of image and video quality.

4. A new approach for predicting the Quality of Experience (QoE) of fish-eye to rectilinear transformed images is proposed. Improved correlation with human opinion is achieved by weighting local quality scores with saliency information.

5. A new automotive specific video quality database is presented consisting of 50 video sequences and associated human saliency data and MOS values. The influence of packet loss on visual QoE for automotive data networks is investigated. The results show that increasing the level of packet loss had almost no effect on visual attention, despite significant differences in the MOS scores of different levels of packet loss.

6. A saliency based framework for No Reference (NR) video quality assessment of packet loss degraded video is proposed. The metric considers the visibility of packet losses in automotive scenes. It is computationally efficient and offers improved correlation with human opinion over existing methods.

1.2.2 Publications

The publications that have resulted from this research are as follows:

Journal Publications (copies of accepted journal papers are included in Appendix C)

- Anthony Winterlich, Ciaran Hughes, Liam Kilmartin, Martin Glavin, Edward Jones; "An oriented gradient based image quality metric for pedestrian detection performance evaluation." *Signal Processing: Image Communication.* Vol. 31, February 2015, Pages 61-75

- Anthony Winterlich, Patrick Denny, Liam Kilmartin, Martin Glavin, Edward Jones; "Performance optimization for pedestrian detection on degraded video using natural scene statistics."*SPIE Journal of Electronic Imaging.* Vol. 23 Issue 6, 2014

- Shane Tuohy, Anthony Winterlich [*], Brian McGinley, Martin Glavin, Edward Jones, Patrick Denny, Liam Kilmartin; "Evaluating the influence of packet loss on visual quality of perception for high bandwidth automotive networks."*Signal Processing: Image Communication.* Vol. 43, April 2016, Pages 15-27

Conference Publications

- Anthony Winterlich, Vladimir Zlokolica, Patrick Denny, Liam Kilmartin, Martin Glavin, Edward Jones; "A saliency weighted no-reference blur metric for the automotive environment." *Proceedings of the Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, 2013

[*] Denotes corresponding author

- Anthony Winterlich, Ciaran Hughes, Liam Kilmartin, Martin Glavin, Edward Jones; "Evaluation of the effect of image artifacts on pedestrian detection in the automotive environment." Transport Research Arena (TRA) Conference, Athens, April 2012, Gold medal winner in the Young European Arena of Research competition (field of Safety and Security).

1.3 Thesis Structure

The remainder of this thesis is organised as follows:

Chapter 2 - Background

This chapter describes the overall process of digital image creation that leads to an "uncompressed" image and discusses the automotive imaging context in which images gathered on multiple cameras are compressed and transmitted over a local network. It also provides background on the state of the art in current objective image quality research. Existing full-reference (FR) and no-reference (NR) image quality metrics (largely applied in consumer applications) are described and their relative merits and disadvantages are discussed. The state of the art in pedestrian detection algorithms is also introduced. In particular, the pedestrian detection algorithms used in this research are discussed in detail.

Chapter 3 - Full reference Image Quality Assessment for Pedestrian Detection

This chapter evaluates the effects of video acquisition and compression artifacts on the performance of a number of state-of-the-art pedestrian detection algorithms. It is demonstrated that the HVS may not perceive distortions that adversely affect machine vision performance. As a result, existing FR image quality metrics are not necessarily accurate predictors of machine vision performance on automotive video sequences. To address this problem, a novel, computationally inexpensive, FR objective quality metric based on HOG vectors is proposed for the automotive environment.

Chapter 4 - Detection Optimization using Natural Scene Statistics

In this chapter Natural Scene Statistics (NSS) are used to blindly categorise distorted video frames by distortion type and level without the use of an explicit reference. Image quality statistics are combined with a multi-classifier detection framework for optimal pedestrian detection performance with varying image quality. The proposed method provides statistically significant improvements over current approaches based on single classifiers when tested on two large pedestrian databases containing a wide variety of artificially added distortion. The improvement in detection performance is further demonstrated on real video data captured from multiple cameras containing varying levels of sensor noise and compression artifacts.

Chapter 5 - Evaluating the Influence of Saliency on Perceptual Quality in Automotive Vision Systems

This chapter examines the issue of subjective quality in automotive video. First, the results of a subjective image quality evaluation specifically of automotive images are presented. Based on this subjective test, a new approach for predicting the Quality of Experience (QoE) of fish-eye to rectilinear transformed images used in automotive vision applications is proposed. Fundamental descriptors from the Fourier transform are used to predict perceptual quality. It is demonstrated that locally weighting descriptors according to visual saliency maps improves correlation with subjective MOS. Additionally, an automotive specific video quality database was gathered for this research, consisting of 50 video sequences and associated human saliency data and MOS values. The influence of packet loss on visual QoE for high bandwidth automotive networks is examined, using this new database. The results show that increasing the level of packet loss has almost no effect on visual attention, despite significant differences in MOS values with different levels of packet loss.

Chapter 6 - Conclusions and Future Work

This final chapter revisits the work presented throughout this thesis and summarises the main results and conclusions reached. Some potential avenues for future work are also briefly explored.

Chapter 2

Background

2.1 Introduction

This chapter describes the overall process of digital image creation and discusses the state of the art related to the research described in this thesis, as evidenced by relevant literature. A discussion on current approaches to image quality assessment is presented. Relevant IQA methods are described and their potential advantages and limitations discussed in the context of automotive visual quality assessment.

2.2 Digital Image Creation

2.2.1 Image Optics

In a typical digital camera the lens forms an image of the scene on the digital image sensor. Anti-aliasing and infrared cut off filters are situated between the lens and sensor to prevent unwanted spatial and spectral scene components from being imaged, and a cover glass is placed over the lens to protect the imaging surface from dust. The basic imaging element of a digital camera sensor is called a pixel. The optical efficiency of a digital camera is determined by the size of the photosensitive area of each pixel and by the point spread function (PSF) of the lens. The PSF is the image created by the lens of a point of light in the scene. In general the higher the quality of the lens, the narrower the PSF. In order to maximise optical efficiency the PSF should be sufficiently small so that the image of a point of light falls entirely on the photosensitive area of a single pixel. In many consumer cameras this desirable property of optical imaging is too costly to realise. In practice, the PSF of the lens can be larger than the photosensitive area of the pixel, resulting in decreased optical efficiency. Light may also leak into neighbouring pixels to create a phenomenon know as cross-talk. A more detailed discussion of the PSF and the pixel sizes of modern cameras can be found in [24].

2.2.2 Image Filters

Since a digital sensor consists of a rectangular grid of pixels, it captures a sampled version of the image formed by the lens. In signal and image processing, aliasing is an effect that causes different signals to become indistinguishable when sampled. It causes distortions or image artifacts that result in the image reconstructed from samples being different from the original image. Aliasing in an image is usually associated with distortions in high frequency regions and can be reduced by reducing the pixel size, however this approach is expensive as it requires additional processing and storage resources and also necessitates a reduction in size of the PSF of the lens.

Instead, aliasing is typically reduced by band limiting the captured image through the use of an anti-aliasing filter, wherein the image is optically low pass filtered, thus eliminating the high frequencies that cause aliasing.

For low-end consumer image optics the cost of including an anti-aliasing filter can be prohibitive and so an alternative is to use a lens of lower quality, which produces a more blurred image. Although such a design choice produces a fixed upper bound on image quality, the cost-quality trade-off is often acceptable for consumer applications. The principle of anti-aliasing remains the same, namely eliminating the higher spatial frequencies from the image formed on the sensor.

2.2.3 Sensor Optics

Currently, all image sensors in digital cameras use integrating sensing technology. Each pixel accumulates light-induced photo-charge for a finite exposure time and is then read out. Most sensors convert 20 - 60% of photons into charge. The signal output from the sensors is typically linear with accumulated charge. Sensors are usually grouped according to their fabrication process into Charge-Coupled Devices (CCD) or Complementary Metal Oxide Semiconductor (CMOS) sensors. The key difference between these groups is that CCD sensors transport charge from each pixel to an output gate, where charge is converted into a measurable signal. CMOS sensors convert charge into voltage within each pixel and send that signal to the sensor output. In the automotive industry, due to their lower cost and power consumption CMOS sensors are by far the more prevalent.

Sensor artifacts can degrade the quality of the uncompressed image captured by a digital camera in a variety of ways. For example, because there is no light-shielded storage in most CMOS sensors, charge cannot be stored for any significant amount of time before readout, therefore most sensors use a rolling shutter readout scheme, in which pixels are read out one row at a time. Although the rolling shutter approach provides a simple and cost-effective solution for pixel read out, distortions occur because different image lines are exposed over different intervals of time. The motion of either the camera or subject can cause geometric distortions such as skew that can cause problems in automotive imaging.

Defective pixels are another source of sensor artifact. Defective pixels are pixels whose response is abnormal enough to be disregarded after readout. Pixel defects may arise from an impurity in the silicon crystal, a flaw in a filter, an electronic fault or a surface flaw on the sensor such as a scratch or a piece of dirt. Isolated pixel defects are common and can be easily concealed in the processing path, however because CMOS sensors address pixels with row and column selection, both row and column defects are possible. Furthermore, defects on the surface of the sensor may affect a cluster of pixels. In some CMOS designs, multiple pixels (usually 2 or 4) share the same circuitry for conversion of charge to voltage, so a failure in circuitry can affect multiple pixels.

Camera noise is a further source of image defects that arises in digital cameras. An example is “fixed pattern noise”, which is characterised by the same pattern of brighter and darker pixels occurring in images taken under the same illumination conditions. It occurs due to variations in the geometries and sizes of individual pixels and can be mitigated to a large extent by camera calibration under known illumination conditions. Other noise sources occur from the electronics and tend to be more noticeable in low light conditions when pixel gain is increased.

Chromatic aberrations occur in the image capture process due to the nature of the lens material and the varying effects that it has on different wavelengths of light. In the final image, a common result is lateral chromatic aberration, where the colour channels are displaced on the sensor, or axial chromatic aberration, where the colour channels have different sharpness [25].

Notwithstanding the many image defects that can occur in the image acquisition process, this thesis is primarily concerned with distortion effects that occur in the post processing pipeline after images are acquired in the sensor (including in particular compression artifacts and transmission errors); therefore, the effects of distortions that occur during acquisition will not be considered in further detail.

2.2.4 Digital Image Formats

The Bayer pattern, introduced in 1976 is a colour filter array designed to mimic the light sensitive physiology of the HVS. The HVS obtains luminance information mostly from green wavelengths of light, while colour information is derived from blue and red wavelengths. Thus the Bayer pattern has twice as many green sensors as red or blue. The resulting pattern, illustrated in Figure 2.1 is used almost exclusively on modern image sensors.

From Figure 2.1 it can be seen that colour pixels of a Bayer pattern sensor do not overlap each other spatially. This Bayer pattern image that is obtained from the image sensor is referred to as a raw or uncompressed image. The raw image is

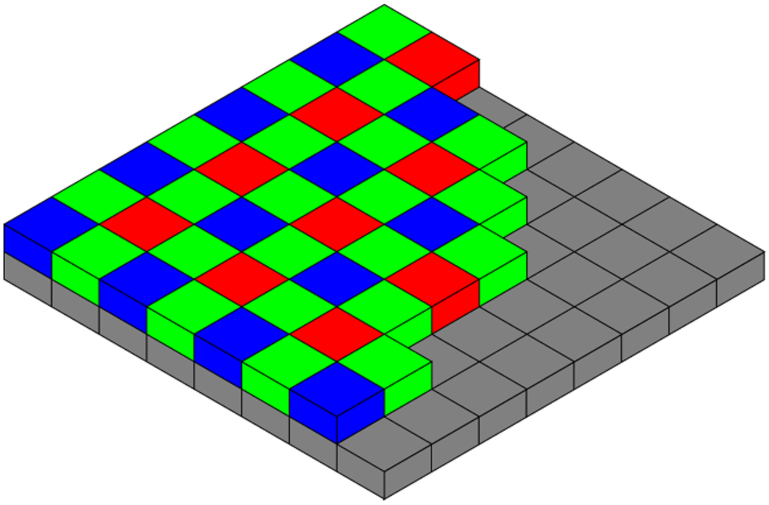


Figure 2.1: Bayer colour filter array (figure taken from wikimedia.org, reproduced under the GNU free documentation license).

converted to other digital image formats for further post processing.

The next step in the digital image pipeline is typically a conversion of the image to the RGB colour space to allow for gamma conversion. RGB is a convenient colour model for computer graphics because the HVS works in a similar, though not identical way, to an RGB colour space. However, standard RGB colour space formats use a non-linear encoding of the intended intensities of the primary colours, which is further dependent on the luminance and tonal distributions of the current scene. As this is too complex to correct in a single step, the gamma is typically corrected at this stage with colour and luminance balancing occurring after YC_bC_r conversion. YC_bC_r is used as a part of the color image pipeline in video and digital photography systems. Y is the luma component and C_b and C_r are the blue-difference and red-difference chroma components. YC_bC_r is not an absolute colour space but rather it is a way of encoding RGB information that takes advantage of a particular characteristic of human perception. The HVS is less sensitive to the compression of colour data than to the compression of luminance data. Moving the image to YC_bC_r format is a first step towards image compression. The most common YC_bC_r format is YUV. The YUV model defines a colour space in terms of one luminance (Y) and two chrominance (UV) components.

2.2.5 Image Compression

JPEG: In the automotive industry JPEG is the most common type of image compression [26, 27]. The term JPEG is an acronym for the "Joint Photographic Experts Group", who are responsible for the JPEG standard. The starting point for JPEG compression is the corrected YUV image. The two colour components are typically down-sampled since they are of less importance to the HVS. Next, the three channels, Y, U and V are split into 8×8 macroblocks. Each 8×8 block is then transformed into the frequency domain via a normalized, 2D discrete cosine transform (DCT). The next step in the compression process is quantization. The HVS is not good at detecting high frequency variations in brightness, therefore the amount of high frequency information in the image can be greatly reduced. A simple approach is to divide each frequency domain component by a fixed value and round to the nearest integer. The rounding operation, together with chroma sub-sampling are the *lossy* parts of the JPEG compression algorithm. Typically, many high frequency components of the DCT domain will be rounded to zero. Entropy encoding is then employed to further compress the image, however this process is *lossless* and therefore does not affect the image quality.

JPEG 2000: JPEG 2000 (JP2) is an image compression standard and coding system created by the Joint Photographic Experts Group committee in 2000 with the intention of superseding the JPEG standard. After color transformation, the image is split into so-called tiles, rectangular regions of the image that are transformed and encoded separately. Tiles can be any size, and it is also possible to consider the whole image as one single tile. Dividing the image into tiles is advantageous since the decoder needs less memory to decode the image and can opt to decode only selected tiles to achieve a partial decoding of the image. However a disadvantage of this approach is that the quality of the picture decreases due to a lower peak signal-to-noise ratio and using many tiles can create a blocking effect similar to the JPEG standard.

The tiles are then wavelet transformed to an arbitrary depth. After the wavelet transform, the coefficients are scalar-quantized to reduce the number of bits used to represent them. The output is a set of integer numbers that are encoded bit-by-bit. The parameter that can be changed to set the final quality is the quantization step size: the greater the step size, the greater the compression and resulting loss of quality.

In the automotive applications considered in this thesis, multiple camera streams may be employed for advanced driver assistance systems (ADAS) such as automatic breaking and reverse parking assistance. As a result, images or video from multiple sources must be compressed and transmitted over an in-car network before either being displayed to the driver or used for computer vision applications. Hence, this thesis deals with the image quality degradations associated with image compression and network transmission in the automotive environment, from the point of view of the driver's perception of quality, and also examines the impact of image degradations in automotive computer vision applications.

2.3 Objective Metrics for Perceptual Quality

In general, there are two categories of IQA methods, namely subjective and objective assessment. As mentioned in Chapter 1, the Mean Opinion Score (MOS) is the most widely used subjective assessment technique, however there are several problems with obtaining MOS scores for image analysis. A wide variety of possible test methods and test parameters must be considered and meticulous set-up and control of each experiment is required. Furthermore a large number of observers are required and their subjective assessment of quality must be screened. The process of subjective testing is complex and time consuming and the results from such tests are generally useful only for development purposes; clearly such tests cannot be used for production testing, in-system debugging, or be incorporated into real-time vision systems.

To facilitate automatic assessment of image and video quality many objective IQA algorithms have been proposed in the literature. These algorithms can be classified according to their use, or not, of an original reference image. "Full-reference" (FR) IQA algorithms are algorithms that make use of a reference image. The reference image is assumed to be of perfect visual quality and is usually compared with the distorted image at the pixel level. Differences between the images are quantified in order to derive a quality score. A number of so-called "reduced-reference" (RR) image quality metrics have also been proposed, removing the need to store the entire original image by computing a few statistics from the distorted image and comparing them with the corresponding stored statistics of the original image. “No-Reference” (NR) IQA algorithms, on the other hand attempt to rate quality without the use of a reference image.

2.3.1 Full Reference Image Quality Algorithms

Full reference (FR) image quality metrics are important tools for optimizing system design parameters associated with image acquisition, compression and transmission. In this section some FR perceptual quality algorithms that are commonly cited in the literature are reviewed. A number of these are based on relatively simple measures computed from the reference and test images, while others are more sophisticated and utilise known elements of the human visual system (HVS). In general, as noted previously, the goal of these algorithms is to produce a metric for image quality that correlates well with subjective opinions based on MOS.

PSNR: For over half a century the peak signal to noise ratio (PSNR) has been the most widely used performance metric in the field of image processing [28]. It is defined as follows: Let $x = \{x_i | i = 1, 2, \ldots, N\}$ and $y = \{y_i | i = 1, 2, \ldots, N\}$ represent two images, where N is the number of pixels and x_i and y_i are the intensities of the i^{th} pixels in images x and y respectively. Then

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (x_i - y_i)^2 \tag{2.1}$$

and

$$PSNR = 10Log_{10} \frac{L^2}{MSE} \tag{2.2}$$

where L is the dynamic range of allowable pixel intensities. A disadvantage of PSNR is that it assumes that distortion is only caused by additive signal-independent noise. Unfortunately, this assumption is a gross over-simplification as modern image processing techniques such as compression and distortion correction can introduce degradation in images in a variety of ways. As a result, correlation between PSNR and subjective image quality is known to be poor [29]. Nevertheless, PSNR remains a widely used image quality metric since it is easy to compute. It has a clear physical meaning, i.e. it represents the energy of the error signal and it is preserved after unitary transformations such as the Fourier transform. Additionally, PSNR is often used as a benchmark for more advanced IQA metrics. However, in recent years, much more effort has been devoted to assessing the perceptual quality of an image or video sequence [30] and many alternative IQA metrics have been proposed.

VSNR: Chandler and Hemami presented a wavelet-based visual signal to noise ratio (VSNR) measure for natural images in [31]. VSNR operates under the following framework: firstly, the "visibility" of the distortions in the image is determined. This is achieved through the computation of salient distortion thresholds which are derived from wavelet based models of visual masking and visual summation. Distortions which are below the saliency threshold are considered non-salient, that is, they are not considered to affect the visual quality of an image. Hence images containing only non-salient distortions are considered to be of pristine visual quality. If distortions are above the saliency threshold, a multi-scale wavelet decomposition of the image is performed. The VSNR measure is given by:

$$VSNR = 10log_{10}\Big(\frac{C^2(I)}{VD^2}\Big) \tag{2.3}$$

where $C(I)$ denotes the RMS contrast of the original image I and VD is a measure of visual distortion derived from a linear combination of perceived distortion and perceived global precedence. In tests on the Laboratory for Image and Video Engineering (LIVE) image database [32], VSNR correlates well with subjective scores, particularly with additive white Gaussian noise (AWGN) distortion where correlation with subjective opinion is 0.978 using Pearson's linear correlation coefficient. A limitation of VSNR is that it is measured on the entire image and hence does not offer spatially localized quality information that could be useful in block-based image processing algorithms.

NQM: The noise quality measure (NQM) was originally proposed to evaluate the quality of images degraded only by noise [33]. Nevertheless, NQM also shows acceptable results in the presence of other types of image degradation [34]. The NQM algorithm operates by processing the original and distorted images through a model restoration algorithm based on Peli's contrast pyramid [35]. The authors modify the pyramid by defining a threshold that varies for each spatial frequency band and each pixel in the bandpass images, to account for contrast masking. If $O_s(x,y)$ and $I_s(x,y)$ denote the processed reference and processed distorted images respectively then the NQM is given by:

$$NQM_{db} = 10log\frac{\Sigma_x\Sigma_y O_s^2(x,y)}{\Sigma_x\Sigma_y(O_s(x,y)-I_s(x,y))^2} \tag{2.4}$$

An advantage of this approach is that variations in contrast sensitivity with distance and image dimensions can be taken into account. Like VSNR, NQM correlates well with subjective quality scores for AWGN degraded images, however, in general, more recent metrics offer statistically superior correlation with subjective mean opinion scores [36].

While the above image quality metrics rely on low-level properties of vision, alternative approaches attempt to model aspects of the HVS in order to more accurately assess image quality.

SSIM: In [15] Wang et al. proposed the structural similarity (SSIM) index, which is based on the assumption that the HVS has evolved to extract structural information from natural scenes. The perceived quality of an image is therefore related to the structural fidelity between a distorted image and the original. The SSIM system separates the task of similarity measurement into three components: luminance, contrast and structure. First, the luminance of each signal is compared. If x and y are two aligned images, the luminance of x is estimated from the mean intensity, defined as:

$$\mu_x = \frac{1}{n}\sum_{i=1}^{N} x_i \tag{2.5}$$

The luminance comparison function $l(x, y)$ is then a function of μ_x and μ_y. The mean intensity is removed from the signal and the standard deviation is used as an estimate of signal contrast. An unbiased estimate in discrete form is given by:

$$\sigma_x = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_i)^2 \tag{2.6}$$

The contrast comparison $c(x, y)$ is then the comparison of σ_x and σ_y. The signal is then normalized by its own standard deviation, so that the two signals being compared have unit standard deviation. The structure comparison $s(x, y)$ is conducted on the normalized signals $\frac{x-\mu_x}{\sigma_x}$ and $\frac{y-\mu_y}{\sigma_y}$. The correlation (inner product) between these vectors is a simple and effective measure to quantify the structural similarity. Finally the three components are combined to yield an overall similarity measure:

$$S = l(x, y) \cdot c(x, y) \cdot s(x, y)). \tag{2.7}$$

The SSIM index has been shown to correlate highly with human opinion scores of images containing a range of distortions. For example, the algorithm achieved correlation coefficients of 0.970, 0.943 and 0.956 on the AWGN, JPEG and JP2 compression subsets of the LIVE image database, respectively [37].

IFC: Other HVS based models include the Information Fidelity Criterion (IFC) which was proposed in [38]. The IFC is an information-theoretic approach to image quality evaluation in which Gaussian Scale Mixtures (GSMs) are computed in the wavelet domain across multiple sub-bands. Visual quality is then estimated by quantifying the mutual information between the GSM models of the original and distorted images. For the source model a single sub-band of the wavelet decomposition is modeled as a GSM Random Field (RF), $C = \{C_i : i \in I\}$, where I denotes a set of spacial indices for the RF. C is a product of two stationary RFs that are independent of each other:

$$C = S \cdot U = \{S_i \cdot U_i : i \in I\} \tag{2.8}$$

where $S = \{S_i : i \in I\}$ is an RF of positive scalars and $U = \{U_i : i \in I\}$ is a Gaussian scalar RF with mean zero and variance σ_U^2.

The distortion in the channel is modeled as a simple signal attenuation and additive Gaussian noise model in each sub-band:

$$D = GC + V = \{g_i C_i + V_i : i \in I\} \tag{2.9}$$

where C denotes the RF from a sub-band in the reference signal, $D = \{D_i : i \in I\}$

denotes the RF from the corresponding sub-band of the distorted signal, $G = \{g_i : i \in I\}$ is a deterministic scalar attenuation field, and $V = \{V_i : i \in I\}$ is a stationary additive zero-mean Gaussian noise RF with variance σ_V^2. Given the statistical models for the source and distorted channels above, the conditional mutual information, denoted $I\Big(C^N; D^N|S^N\Big)$ is computed for each sub-band. The IFC is then obtained by summing over all sub-bands as follows:

$$IFC = \sum_{k \in sub\text{-}bands} I\Big(C^{N_k,k}; D^{N_k,k}|S^{N_k,k}\Big) \tag{2.10}$$

where $C^{N_k,k}$ denotes N_k coefficients from the RF C^k of the k^{th} sub-band, and similarly for $D^{N_k,k}$ and $S^{N_k,k}$.

VIF: The Visual Information Fidelity (VIF) metric, proposed in [39], has been shown to correlate highly with human opinion for multiple distortion types. VIF is an extension of the IFC, which takes into consideration the fact that the human visual system limits the amount of information that can be extracted from a visual signal. The VIF metric therefore aims to quantify the loss of information to the HVS channel relative to the amount of information lost from the source signal to the distortion channel. Letting $I\Big(C^N; E^N|S^N\Big)$, and $I\Big(C^N; F^N|S^N\Big)$ denote the information that could ideally be extracted by the HVS from a particular sub-band of the reference and distorted images respectively, VIF is given by:

$$VIF = \frac{\sum_{k \in sub\text{-}bands} I\Big(C^{N_k,k}; F^{N_k,k}|S^{N_k,k}\Big)}{\sum_{k \in sub\text{-}bands} I\Big(C^{N_k,k}; E^{N_k,k}|S^{N_k,k}\Big)} \tag{2.11}$$

where, as before, $C^{N_k,k}$ denotes N_k coefficients from the RF C^k of the k^{th} sub-band, and similarly for $E^{N_k,k}$, $F^{N_k,k}$ and $S^{N_k,k}$. A particularly interesting feature of VIF is that linear contrast enhancement of the reference image is taken into account.

Assuming such an enhancement does not add any additional noise, the enhanced image will be rated as superior in quality to the reference image.

MLSIM A Multi-Level Similarity index (MLSIM) for IQA was proposed in [40]. The MLSIM is based on the principle that the HVS determines image quality mainly according to details extracted from low level gradient information. The Prewitt magnitude [41] of the reference f and distorted image g are computed to obtain f_p and g_p. The images are then segmented into multiple levels based on thresholding of their Prewitt magnitudes, to obtain $L_i(f_p)$ and $L_i(g_p)$. For each level i, the coefficients of the first and second order Riesz transforms $R_j(L_i(f_p))$ and $R_j(L_i(f_p))$, $j = 1, \ldots, 5$ are computed. Five different values for the regional mutual information (RMI) proposed by [42] are then computed for each level, with the average RMI value for level i denoted $RMI_i(f, g)$. The MLSIM value is given by:

$$MLSIM(f, g) = \sum_{i=1}^{N} \omega_i \cdot RMI_i(f, g) \tag{2.12}$$

where ω_i, $i = 1, \ldots, N$ are weighting factors for each level.

A disadvantage of the IFC, VIF and MLSIM algorithms is their computation time [43, 44, 40], which limits their utility in real-time applications. More recently, a number of advances in the area of image quality evaluation have been proposed that incorporate sophisticated models of visual saliency information, while also improving computational performance. These include:

FSIM: In [45], Zhang et al. propose a novel feature similarity (FSIM) index for full reference image quality assessment. The main feature used in FSIM is the dimensionless phase congruency (PC) which measures the significance of a local structure. Since PC is contrast invariant, and contrast information affects the HVS's perception of image quality, the image gradient magnitude (GM) is also incorporated into the model as an additional feature. PC and GM play complementary roles in characterizing the local image quality. FSIM has been shown to achieve higher consistency with subjective evaluations than alternative IQA metrics [45]. The metric is computed as follows. Let the similarity between two images f and g be given by:

$$S_L(x) = [S_{PC}(x)] \cdot [S_G(x)] \tag{2.13}$$

where $S_{PC}(x)$ and $S_G(x)$ are similarity measures of the computed PC and GM of the reference and distorted images. Then FSIM is defined as:

$$FSIM = \frac{\sum_{x \in \Omega} S_L(x) \cdot PC_m(x)}{\sum_{x \in \Omega} PC_m(x)} \tag{2.14}$$

where Ω is the entire image spatial domain, and $PC_M(x)$ is the maximum phase congruency of the reference and distorted images.

SR-SIM: The spectral residual based similarity index (SR-SIM) proposed in [44] is a novel, computationally inexpensive image quality metric based on spectral residual visual saliency. SR-SIM differs from FSIM only in that the Spectral Residual Visual Saliency (SRVS), denoted R is used as a substitute for PC so that:

$$SR\text{-}SIM = \frac{\sum_{x \in \Omega} S_L(x) \cdot R_m(x)}{\sum_{x \in \Omega} R_m(x)} \tag{2.15}$$

where Ω is the entire image spatial domain, $R_M(x)$ is the maximum phase congruency of the reference and distorted images, and in this case the similarity measure $S_L(x)$ is given by:

$$S_L(x) = [S_R(x)] \cdot [S_G(x)] \tag{2.16}$$

. Extensive experiments were conducted on three large-scale IQA data sets which indicated that SR-SIM is capable of achieving very high correlation with human perceptual judgement.

IW-SSIM: In [46], Wang and Li weight the local quality measures of a number of full reference image quality metrics with the local information content, which is estimated in units of bits, using advanced statistical models of natural images. The results show that intelligent weighting of local quality scores can considerably improve the correlation of a quality metric to subjective opinion scores. The proposed information content-weighted structural similarity measure (IW-SSIM), is an extension of SSIM. In [40] a multi-scale (MS) image quality approach was proposed that incorporates SSIM scores at different scales. Let $x_{j,i}$ and $y_{j,i}$ be the i^{th} local image patches at the j^{th} scale, and let N_j be the number of evaluation windows in the scale, then the j^{th} scale SSIM evaluation is given by:

$$SSIM_j = \frac{1}{N_j} \sum_i c(x_{j,i}, y_{j,i}) \cdot s(x_{j,i}, y_{j,i})). \tag{2.17}$$

for $j = 1, \ldots, M-1$, and

$$SSIM_j = \frac{1}{N_j} \sum_i l(x_{j,i}, y_{j,i}) \cdot c(x_{j,i}, y_{j,i}) \cdot s(x_{j,i}, y_{j,i})). \tag{2.18}$$

for $j = M$, where M is the number of scales. The overall MS-SSIM measure is then defined as:

$$MS\text{-}SSIM = \prod_{j=1}^{M} (SSIM_j)^{\beta_j} \tag{2.19}$$

where the β_j values are obtained through psychophysical experiments. By combining information content weighting with MS-SSIM, Wang and Li defined an information content weighted SSIM measure (IW-SSIM). Let $w_{j,i}$ be the information content weight computed at the i^{th} spatial location in the j^{th} scale, then the j^{th} scale IW-SSIM measure is defined as:

$$IW\text{-}SSIM_j = \frac{\sum_i w_{j,i} \cdot c(x_{j,i}, y_{j,i}) \cdot s(x_{j,i}, y_{j,i}))}{\sum w_{j,i}}. \tag{2.20}$$

for $j = 1, \ldots, M - 1$, and

$$IW\text{-}SSIM_j = \frac{1}{N_j} \sum_i l(x_{j,i}, y_{j,i}) \cdot c(x_{j,i}, y_{j,i}) \cdot s(x_{j,i}, y_{j,i})). \tag{2.21}$$

for $j = M$. The final IW-SSIM metric is thus computed as:

$$IW\text{-}SSIM = \prod_{j=1}^{M} (IW\text{-}SSIM_j)^{\beta_j} \tag{2.22}$$

2.3.2 Reduced Reference Image Quality Algorithms

Reduced-reference (RR) image quality metrics provide a solution that lies between FR and NR models. They are designed to predict the perceptual quality of distorted images using only partial information from the reference images. These partial RR features usually have a much lower data rate than the image data. RR methods are useful in a number of applications, for example in real-time visual communication systems, as they can be used to monitor image quality degradations and control the resources available for video streaming; the partial information about the reference image can be communicated along with the distorted image with relatively low overhead. Existing RR IQA algorithms typically use one of three different approaches [47]. The most common approach is based on modelling image distortions. This approach is particularly useful for specific application environments where the distortion type is known. For example, in [48] Gunawan and Ghanbari presented a reduced reference objective quality metric for compressed video. Discriminative analysis of harmonic strength was computed from edge-detected pictures to provide harmonic gain and loss information. Harmonic gain and loss correspond to blockiness and blurring artifacts associated with image compression. The proposed approach achieved good correlation with subjective opinions from the Video Quality Experts Group (VQEG) Test Phase 1 video sequences [49]. In [50], a RR model was described wherein features were extracted from spatial-temporal blocks to determine quality. The first feature, a measure of overall spatial information is used to detect localized blurring in the distorted video sequence, while a second feature, which measures the angular distribution of spatial gradients, provides a simple means of incorporating variations in the sensitivity of the HVS to angular orientation. A third feature is used to measure distortions in the chrominance channels. These features have been shown to provide close correlation to MOS values and require little additional bandwidth [50, 51].

The second type of approach is based on modelling the HVS e.g. [52, 53]. Typically, perceptual features motivated from computational models of low level vision are extracted from the reference image to provide a reduced description of image quality. An advantage of this approach is that the perceptual features extracted from the image are not directly related to any specific distortion type and hence could potentially provide distortion independent assessment of image quality. A third approach is based on modelling Natural Scene Statistics (NSS). The underlying assumption behind NSS approaches is that most image distortions disturb image statistics and make the distorted image somewhat "unnatural". The distance between the reference and distorted image statistics can thus be measured and used to predict degradations in perceptual image quality. In [54], Wang and Simoncelli proposed an RR

IQA method based on a NSS model in the wavelet transform domain. Specifically, it was shown that the marginal distribution of the wavelet coefficients of a particular sub-band changes in different ways for different types of image distortions. The Kullback-Leibler (KL) distance is used to measure the difference in marginal probability distributions of the wavelet coefficients extracted from both reference and degraded images. An advantage of this approach is that only a relatively small number of RR features are required for image quality evaluation. A similar approach is used in [47], wherein Li and Wang describe a RR IQA method that is inspired by the divisive normalisation transform (DNT) described in [55]. By using a GSM statistical model of image wavelet coefficients, DNT transforms of the reference and degraded images are computed. Again, quality is assessed by comparing the difference in probability distributions of extracted features using the KL distance.

A family of RR video quality metrics was introduced in [56] and [57] that incorporate both spatial and temporal differences in entropy. A GSM model for the wavelet coefficients of frames and frame differences was used to measure the amount of spatial and temporal information differences between the reference and degraded videos respectively. The spatial and temporal differences were then combined to obtain spatio-temporal reduced reference entropic differences.

Finally, a recently proposed RR quality metric for compressed video [58] combined two of the above approaches by incorporating spatial information loss and the temporal characteristics of the inter-frame histogram. In the spatial domain, an energy variation descriptor is proposed to measure the energy change of each individual encoded frame after quantisation. In the temporal domain, the Generalised Gaussian

Density (GGD) function is used to capture the statistics of the interframe histogram distribution. The proposed metric outperformed FR metrics such as PSNR and SSIM in an evaluation on the LIVE video database.

2.3.3 No-Reference Image Quality Algorithms

While the availability of a reference image greatly simplifies the task of quality assessment, in automotive vision applications such algorithms are generally limited to use in the system design stage where they may be used to optimize visual quality and machine vision performance. However, for real-time applications, a reference image is typically unavailable when image quality needs to be computed. “No-reference” (NR) image and video quality algorithms, on the other hand, have the potential to be incorporated into real-time automotive vision systems where they may be used, for example, to dynamically control compression rates in compressed video in order to guarantee a particular quality of service.

Like RR metrics, the majority of NR algorithms are distortion specific; for example, there are many algorithms that rate the quality of compressed [59, 60, 61, 62, 63, 64, 65], blurred [66, 67, 68, 23, 69] or channel distorted [70, 71] images. These NR algorithms estimate the level of a particular distortion present in an image and map this value to a quality score by using *a priori* information on subjective opinion scores. Distortion specific metrics are useful in systems where there is a single predominant distortion, for example in JPEG2000 compression, in which “ringing” artifacts are prevalent [72, 73]. Ringing is caused by the quantisation of high frequency coefficients in transform coding and is characterized by ripples around sharp edges. In [61] a NR algorithm for ringing artifacts was developed. The algorithm estimates the visibility of ringing artifacts by comparing them to the activity of the local background, and was shown to correlate closely with subjective data.

While distortion-specific RR IQA algorithms produce a close correlation to human opinion for specific distortions, in the automotive environment such algorithms may be somewhat limited, since degradations in video quality can occur from various other sources such as sensor noise, transmission artifacts or video compression. Furthermore, automotive vision systems operate over a diverse range of landscapes and environmental conditions making it difficult to predict the most prevalent distortion, which will likely change depending on scene complexity and illumination [21].

Recently, distortion-independent approaches which are based on the statistical properties of natural images have been shown to provide good performance in predicting perceived image quality.

Such approaches are based on the hypothesis that natural images follow regular statistical properties that are altered by the presence of distortions [74]. As an example, consider the grey-scale values of two neighbouring pixels. If many different locations in an image are selected in random order, and the grey-scale values of the pixels as the observed values of two random variables are considered, these random variables will not be independent. Intuitively, it is clear that two neighbouring pixels tend to have very similar grey-scale values. Such regular properties associated with natural images are referred to as NSS. Deviations from these statistics can be quantified to predict perceptual image quality. Since NSS models effectively assess the "naturalness" of an image, they necessarily provide a distortion independent assessment of perceived image quality. However, research has shown that different distortion types alter the statistics of natural images in characteristic ways, and hence NSS have been used to categorise degraded images by distortion type for example in [75] and [76].

In [77], the Distortion Identification-based Image Verity and INtegrity Evaluation (DIIVINE) index was proposed. The algorithm is based on a two-stage framework involving distortion identification followed by distortion-specific quality assessment. A distorted image is decomposed using a scale space orientation decomposition to form oriented band pass responses. A series of statistical features are then extracted from the sub-band coefficients and stacked to form a vector, which is a statistical description of the distortions in the image. The feature vectors can then be utilized to evaluate the probability of the image containing a particular distortion.

A similar approach called BLind Image Integrity Notator using DCT Statistics (BLIINDS-II) was described in [78] and [79]. A small number of features were computed from an NSS model of block DCT coefficients. These features were then used to train a regression model which accurately predicted perceived image quality. While both of these NR IQA metrics provide high correlation with perceptual image quality, computation of the required features is expensive [75] and hence achieving real time implementation would likely be challenging. To this end, transform-free models for NR IQA have been developed, for example, a general-purpose NR IQA approach based on visual codebooks was proposed in [80] which utilized Gabor-filter-based local features extracted from local image patches to capture NSS. Two relatively new image quality metrics, Blind/Referenceless Image Spatial QUality Evaluator (BRISQUE) and Natural Image Quality Evaluator (NIQE) were introduced in [81] and [75]. Both the BRISQUE and NIQE metrics have been shown to offer comparable performance with the transform-based NR IQA models mentioned above. However both algorithms operate on multiscale spatial pixel data and hence are inexpensive to compute, making them good choices for real-time automotive applications. BRISQUE uses "quality-aware" spatial features to train a regression model for IQA, while NIQE develops a model for undistorted "pristine" images and measures deviations in the statistics of the test image from those in the pristine model. Finally, in [82] Zhang and Chandler proposed an efficient NR IQA algorithm using a log-derivative statistical model of natural scenes. The method is termed DErivative Statistics-based QUality Evaluator (DESIQUE) and utilizes statistical features from both the spatial and frequency domains. Although similar to BRISQUE, the log-derivative based analysis used in DESIQUE provides greater sensitivity to local contrast changes and allows for easier modelling of extracted features. These improvements mean that DESIQUE achieves statistically significant performance improvements over other transform-based NR IQA algorithms, but at much greater computational efficiency [82].

2.4 Objective Quality Assessment for Machine Vision

In the context of advanced driver assistance systems, numerous machine vision algorithms have been developed for tasks including object detection. One such task of particular interest is automated pedestrian detection. In recent years, much research has been devoted to this subject, with numerous algorithms having been proposed in the literature [83, 84, 85, 86]. Although computer vision approaches have made significant progress in this area, there is still room for improvement, particularly for applications that require very accurate responses in real time. A key challenge lies in developing algorithms that exhibit high accuracy and reliability across a wide variety of environmental, implementation and inter-vehicle communication factors, such as illumination, scene content, capture geometry and image compression. A limited body of work has examined various factors that affect detection performance, for example, [87] examines the impact of pedestrian size and position, as well as occlusion statistics on pedestrian detection performance, while environmental and illumination conditions are discussed in [88]. Certainly these factors affect image quality and hence detection performance. However, there is no clear distinction between scene dependent quality factors such as occlusion and illumination, and image quality factors associated with image capture and compression. In this thesis image quality factors associated with image capture, lossy compression and transmission errors are considered, all of which are important in automotive vision systems. Quantifying the loss of performance in detection algorithms due to these degradations is of great interest to the automotive vision community, since knowledge of detection performance in the presence of distortions could be used to guide system configuration or control of compression ratio for driver assistance systems, ensuring good overall system design, and robust detection performance across a wide range of image and video quality. To date, relatively little research has been carried out in this area. Examples in an alternative application area include [89], in which the authors examined the influence of image distortions on a commonly used face detection algorithm, and [90], wherein acceptable bit-rates for human face identification were examined. In [91], the influence of image quality degradations from JPEG and H.264 compression algorithms on the performance of an infra-red pedestrian detection algorithm were investigated, however results were reported only in terms of tracking rate at 4 seconds before impact, and the maximum distance at which a pedestrian can be detected. These quality criteria are insufficient to describe detection algorithm performance since there is no consideration of false positives.

2.5 Concluding Remarks

This chapter has described the overall process of digital image creation that leads to an "uncompressed" image. The state of the art in current approaches toward image quality assessment has also been discussed. Considering the many automotive imaging applications that have been deployed in recent years, there is a surprising lack of analysis of image quality in the automotive environment. While some research has been carried out on factors that influence machine vision performance, little effort has been focused towards deriving quality parameters specifically for automotive applications. This thesis aims to address this gap by examining the influence of image degradations on automotive systems and assessing the suitability of current image quality metrics that could be used in this space.

The next chapter discusses the effects of some transmission artifacts on the performance of a number of common pedestrian detection algorithms, and develops an image quality metric that is designed to accurately predict pedestrian detection algorithm performance in the presence of such artifacts.

Chapter 3

Full Reference Image Quality Assessment for Pedestrian Detection

3.1 Introduction

In typical automotive vision systems, degradation in image quality can occur for a number of reasons. Among the more important sources of degradation are lossy compression of images or video, or noise originating in the image acquisition or transmission process (e.g. sensor noise). Quantifying the loss of performance of detection algorithms due to these degradations typically requires the availability of a "gold" standard in the form of annotated test databases which are both expensive and time consuming to produce. It would be desirable to utilise an objective IQA metric that accurately predicts the performance of machine vision algorithms on degraded images and video themselves. Such a metric could be used as a predictor of the impact of choice of system parameters (e.g. compression ratio (CR) in image/video compression) on pedestrian detection performance. It may also obviate, or at least reduce, the need to carry out extensive machine vision performance evaluation using large annotated data sets.

In this chapter the effects of transmission artifacts on the performance of a number of state-of-the-art pedestrian detection algorithms are evaluated. It is demonstrated that the HVS may not perceive distortions that adversely affect machine vision performance; consequently, existing FR image quality metrics (which correlate with HVS performance) are not necessarily accurate predictors of machine vision performance on transmitted video sequences. To address this problem, a novel, computationally inexpensive, FR objective quality metric based on histograms of oriented gradients is proposed. The proposed metric is specifically designed to predict detection algorithm performance in the presence of degradations. The metric can be used at the system design stage in order to optimize image capture parameters for machine vision performance without the need for annotated test databases.

3.2 Pedestrian Detection Algorithms

The field of pedestrian detection research has been extremely active in recent years with many implementations of pedestrian detection algorithms proposed in the literature. For a comprehensive evaluation of the state-of-the-art in pedestrian detection algorithms, the reader is referred to [87]. In this section we briefly outline some of the key features extracted from an image that are useful for object detection, and that form the basis for the work described in this chapter.

Many of the algorithms described in recent literature utilise some form of Histogram of Oriented Gradient (HOG) descriptors, first proposed in [85]. HOGs are feature descriptors used for the purpose of object detection in which local object appearance is characterized by the distribution of local intensity gradients or edge directions. Shape features are another common cue for detection. Gavrila and Philomin [92] employed the Hausdorff distance transform and a template hierarchy to match image edges to a set of shape templates, while Wu and Nevatia [93] used a large pool of short line and curve segments known as 'edgelets' to represent shape locally. Another feature used for pedestrian detection are 'shapelets' [94] which are shape descriptors discriminatively learned from gradients in local patches.

More recent algorithms look to improve pedestrian detection performance by exploiting a combination of features. Wojek and Schiele [95] showed that a combination of Haar-like features, shape context [96] and HOG features outperforms any individual feature. Wang *et al.* [97] combined a texture descriptor based on local binary patterns (LBP) [98] with HOG, while Felzenszwalb et al. [99] described a framework including the detection of object parts and a statistically learned deformable model that relates these parts. The Pairwise Geometrical Histogram (PGH) is a generalization of the Chain Code Histogram (CCH). It is a powerful shape descriptor that is applied to contours matching and is not affected by rotation. Recently, in [100], Yong et al. combined Haar-like features and PGHs for vehicle detection, while Yao and Deng [101] combined shapelet and Haar-like wavelets to develop a robust pedestrian detection approach.

Meanwhile, work on improving the computational efficiency of feature detection includes Zhu et al. [102] who exploited the "Integral Histogram" [103] to efficiently compute the HOG feature. In [104], Dóllar et al. proposed a fast method for approximating features at multiple scales using a sparsely sampled image pyramid.

In this chapter we are concerned with evaluating the impact of image degradations on pedestrian detection performance. The performance of three pedestrian detection algorithms on images degraded by distortions commonly found in automotive vision systems is first examined. The three algorithms chosen include the widely used HOG detector [85], as well as two recent state-of-the-art detection algorithms, namely Integral Channel Features (ICF) [86] and Deformable Parts-based Model (DPM) [99]. In the following section each of the algorithms is briefly described and the reasons for choosing the HOG vector as the feature descriptor in the subsequent analysis is explained.

3.3 Histogram of Oriented Gradients

Although research in pedestrian detection is quite diverse, almost all modern algorithms, including the top 14 ranked algorithms assessed in [87], employ some form of HOG vectors. Furthermore, HOG coupled with SVM classification continues to be widely used in automotive applications for different detection tasks, largely due to the fact that recent advances enable real-time in-vehicle implementations [105, 106]. The general idea behind HOG features is that local object appearance and shape can be characterized by the distribution of local intensity gradients or edge directions. Typically, this is achieved by first carrying out image pre-processing such as gamma correction and then dividing the image frame into small spatial regions or ‘cells’. A histogram of edge orientations is computed over the pixels of each cell. The histogram entries are combined to make up the feature representation. For better invariance to illumination, cell values are contrast normalized before processing. This is achieved by grouping cells into larger spatial blocks and contrast normalizing each block. Lowe-style clipped L2 norm (L2-Hys) block normalization, described in [107], has been found to give good performance in subsequent classification tasks. The normalized descriptor blocks are termed HOG descriptors. The detection window consists of a dense, overlapping grid of HOG descriptors and the resulting combined feature vector is commonly processed through a linear Support Vector Machine (SVM) for classification. SVMs are supervised learning models that analyse image features and recognise patterns. Given a set of training examples, each categorised as a positive or negative sample, an SVM training algorithm builds a model that labels new image samples as either positive or negative based on the new image feature's similarity with the training set. A block diagram of the HOG-based pedestrian detection algorithm is presented in Figure 3.1 [85].

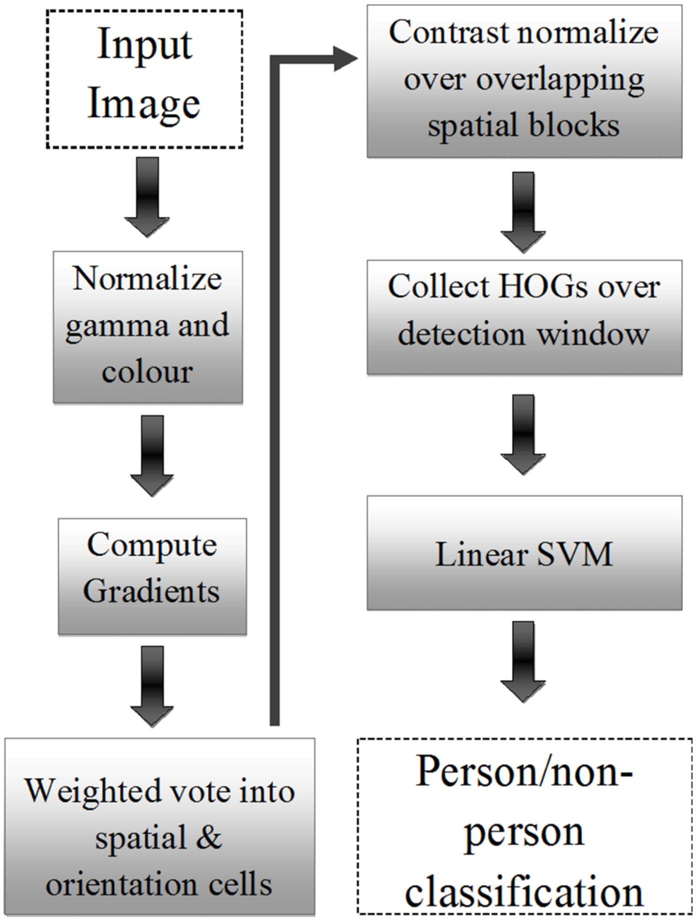

Input Image
Normalize gamma and colour
Compute Gradients
Weighted vote into spatial & orientation cells
Contrast normalize over overlapping spatial blocks
Collect HOGs over detection window
Linear SVM
Person/non-person classification

Figure 3.1: An overview of Dalal and Triggs' feature extraction and object detection chain [85].

The performance of the detection algorithm is affected by the way in which the gradients are computed. Good performance is achieved using a simple 1-D mask with no Gaussian smoothing. HOG vectors have several advantages over other features since they tend to capture edge and gradient structures that are very characteristic of local shape. The HOG feature is also largely invariant to translations or rotations providing they are much smaller than the local spatial and orientation bin sizes. In [85], Dalal and Triggs demonstrated that good parameters for pedestrian detection are found by coarse spatial sampling and fine orientation sampling. Such parameter settings, given in Table 3.1, allow accurate detection even with large movements of pedestrians' limbs. Recent studies [85, 108, 109] have demonstrated that HOG based detectors greatly outperform alternative methods such as the Haar wavelet, Scale Invariant Feature Transformation (SIFT), and shape context approaches.

Table 3.1: Default HOG descriptor Properties

Parameter	**Value**
Window size	64 x 128 pixels
Spatial block size	2 x 2 cells
Cell size	8 x 8 pixels
Number of orientations	9
Overlap	8 x 8 pixels
Gaussian smoothing	No
Histogram normalization	L2-hys
Gamma correction	Yes
Max # of detection scalings	64

Integral Channel Features (ICF) [86] is a detection technique in which multiple registered image channels are computed using linear and non-linear transformations of the input image, and then features such as local sums, histograms, and Haar features and their various generalizations are efficiently computed using integral images. The authors of [86] demonstrated that when designed properly, integral channel features not only outperform other features including HOG, they are also insensitive to exact parameter settings, allow for more accurate spatial localization during detection, and result in fast detectors when coupled with cascade classifiers. The authors reported that the number of frames per second (fps) at which selected channels can be computed for 320×240 images, as tested on a standard PC, were: LUV colour channels at 135 fps, gradient magnitude at 140 fps, and gradient histograms (with 6 bins) at 60 fps. Computing all 10 channel images can be performed at 40 fps. Furthermore, in [110], by efficiently handling different scales and transferring computation from test time to training time, the authors presented a new detector based on ICF which is capable of processing images at over 100 fps.

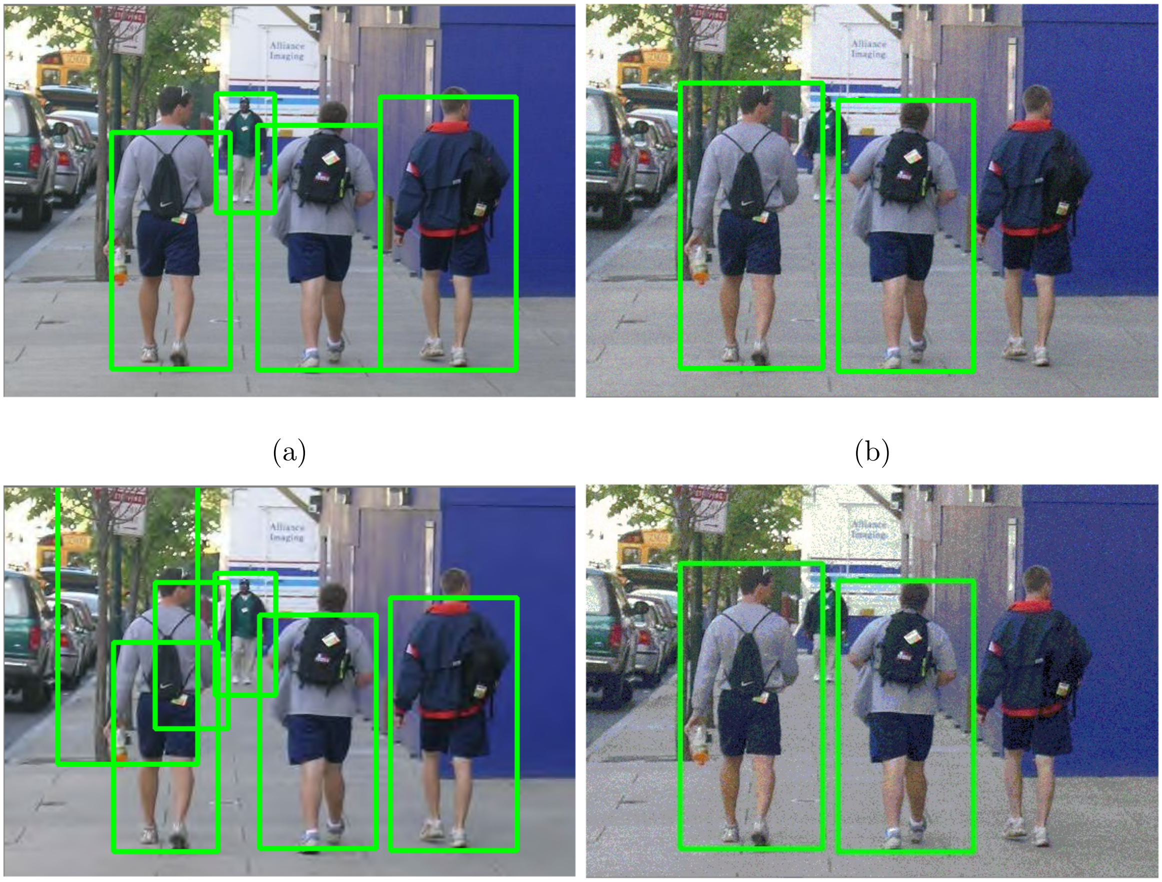
DPM is an object detection algorithm based on mixtures of multi-scale deformable part models. The algorithm, described in [99], is capable of training with partially labelled data and is particularly useful for representing highly variable objects such as pedestrians. The parts based model involves linear filters that are applied to dense feature maps. Feature maps are derived from a variation of the HOG feature. Maps are computed at different scales using a feature pyramid, which is computed by repeated smoothing and sub-sampling of the image. The DPM algorithm achieved very accurate results in the PASCAL object detection challenges [111].

3.4 HOG-based Image Quality Assessment

As indicated in chapter 2, the HVS does not always perceive image degradations that impact the performance of pedestrian detection. It follows that current perceptual IQA algorithms, which correlate closely with human perception, may not be reliable indicators of machine vision performance, which motivates the development of specific IQA algorithms that correlate with pedestrian detection performance. This problem is addressed here. The choice of feature used in the proposed IQA algorithm is motivated by examining some characteristic behaviour of pedestrian detection performance on degraded images.

In general, even a modest amount of AWGN has a significant effect on the performance of detection algorithms. For example, in Figure 3.2, all four pedestrians were detected in the reference image. A failure to detect two pedestrians occurred in the noisy images at PSNR levels of over 30dB (Figure 3.2(b)). Such levels of noise are common in automotive image systems, particularly in low light conditions. On the other hand, the introduction of blocking and ringing artifacts, characteristic of JPEG and JP2 compression algorithms respectively, typically caused an increase in false positives, therefore reducing the precision of the detection algorithm (see Figure 3.2(c)) though without necessarily reducing the true positive rate. Figure 3.2(d) shows an example where the image has been corrupted by AWGN and subsequently compressed with JPEG compression.

Since HOG features and their various generalizations are widely used in pedestrian detection algorithms, examining the HOG vectors of both reference and distorted images can provide a rich set of statistics for image quality evaluation. Figure 3.3 illustrates the HOG vectors extracted from two sample corrupted images (3.3(c), 3.3(d)) and from their corresponding reference image (3.3(b)). The outlines of both pedestrians can be seen quite clearly in the oriented histograms extracted from the reference image and in this case resulted in two correct detections with no false positives. The features extracted from the compressed image clearly show a significant loss in gradient information, but due to the quantization process in JPEG compression, the majority of information loss occurs at high spatial frequencies, which typically represent texture. In this example the majority of loss has occurred on the road and the woman's coat, however the outline of both pedestrians can still be easily



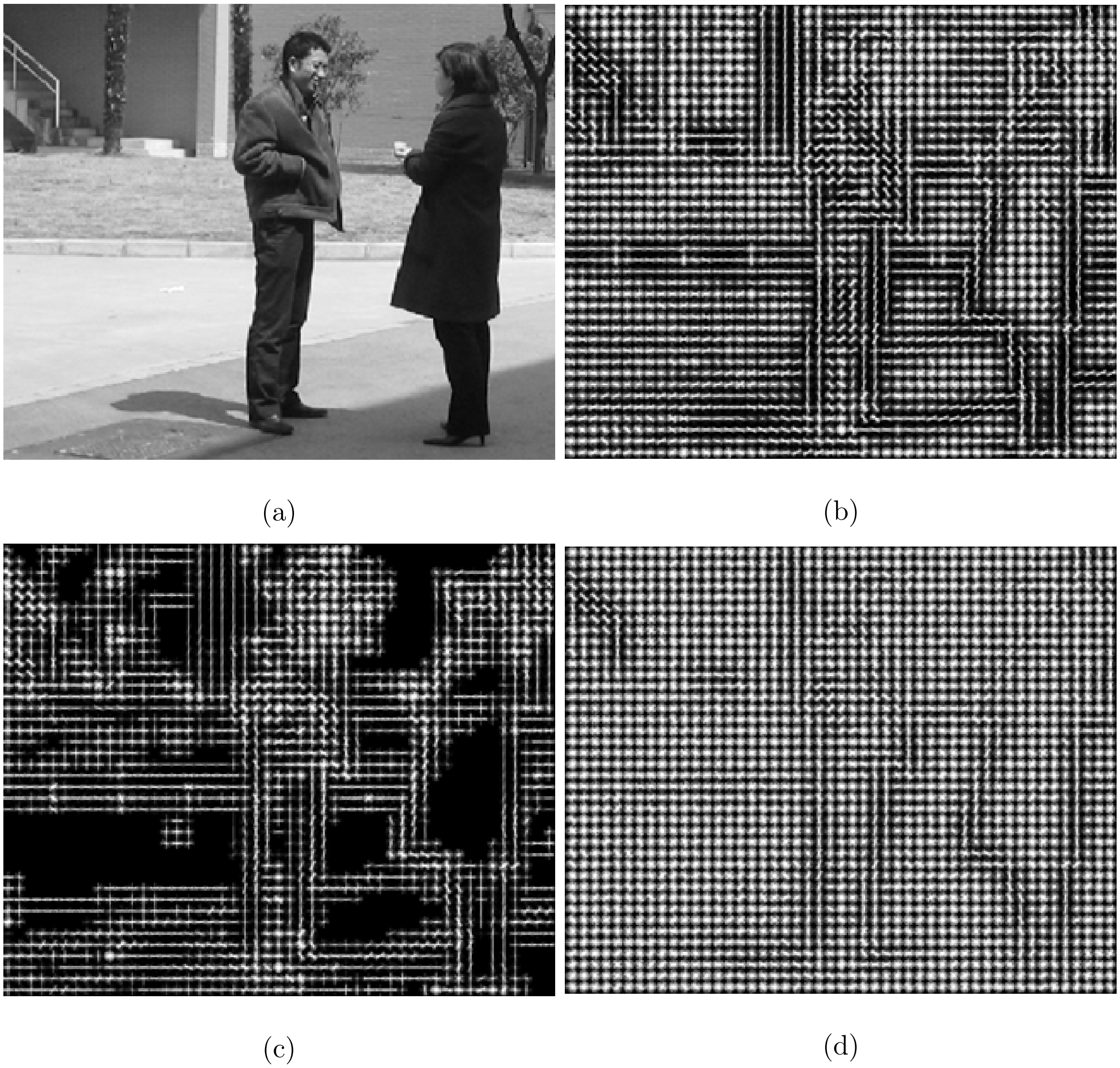
(a) (b)

(c) (d)

Figure 3.2: A typical example of algorithm performance (HOG+SVM) on AWGN corrupted (b) JPEG compressed (c) and both noise corrupted and compressed (d) images.

discerned. This particular example resulted in correct classification of both pedestrians without any false positives. However, in general detection performance tends to deteriorate rapidly at compression ratios over 30:1. This level of compression is common in automotive vision systems, hence balancing the need to compress images with maintaining sufficient quality for machine vision performance is a challenging design problem. The addition of AWGN to the image degrades the extracted features and makes it difficult to determine image structure (Figure 3.3(c)). In this example, neither pedestrian was correctly detected by the detection algorithm.

The individual histogram cells are examined in more detail in Figure 3.4, where a histogram derived from the reference image is shown in 3.4(a). The orientation of the edge is clearly discernible. A close-up view of a single noise degraded histogram cell is shown in Figure 3.4(b). The addition of AWGN leads to a loss of gradient information. Notice that the noise corrupted histogram allocates an almost equal weight to each oriented gradient, making it difficult to discern image edges. This loss of edge information explains the poor performance of the detection algorithm on noise degraded images and also explains why the presence of noise does not contribute to an increase in false positives. On the other hand, reference and corresponding JPEG compressed histogram cells are shown in Figures 3.4(c) and 3.4(d), respectively. Note that the loss of high frequency components in the compressed image leads to a "non-edge" (3.4(c)) being incorrectly classified as an edge. This behaviour explains the increase in the algorithm's false positives as compression rates increase.



(a)

(b)

(c)

(d)

Figure 3.3: HOG vectors associated with (b), an uncompressed; (c), JPEG compressed (with CR=33:1); and (d), AWGN-corrupted (with mean=0 and variance=10^{-2}) version of a reference frame (a).

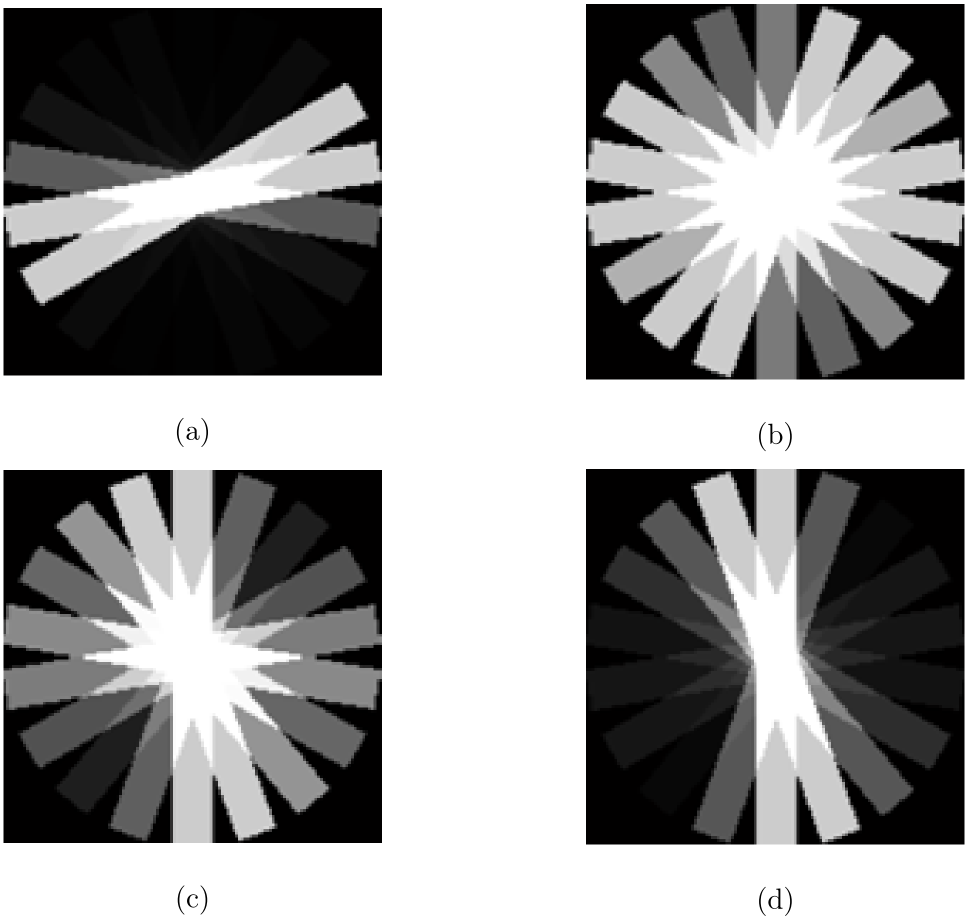


Figure 3.4: HOG vectors associated with (a) a reference "edge"; (b) an AWGN corrupted version of the same reference edge; (c) a reference "non-edge" and (d) a compressed version of (c). The HOGs were computed at a single scale using 8 × 8 pixel cells.

Since the extracted HOG features typically contain the most pertinent structural information in a scene, they have the largest impact on pedestrian detection performance. Changes in gradient information therefore directly influence the final classification result, as outlined in the preceding discussion. Given that the HOG features represent the most relevant information for pedestrian detection, it is hypothesised that an objective metric that captures the error between reference and degraded HOG features will accurately predict the performance of pedestrian detection algorithms on images of varying quality. To investigate this hypothesis, the Mean Squared Error (MSE) is computed between reference and degraded HOG vectors.

There are a number of reasons why the MSE is a good choice for representing the error between both vectors. Apart from being computationally inexpensive it is also a convenient metric for optimization problems since it is differentiable and symmetric. Minimum-MSE problems are generally easy to formulate since the gradient and Hessian matrix of the MSE are easy to compute [30]. In fact, it is often the case that minimum-MSE problems have closed form analytical solutions.

3.4.1 Algorithm Details

The HOG feature is extracted by dividing the image into multiple overlapping blocks of the same size and quantizing the gradient direction of all pixels into 9 orientations. In this analysis, each block is an 8×8 image patch that has 50% overlap with its neighbours. The Matlab code used to extract the HOG vectors is available as part of Piotr's Image and Video Matlab toolbox [112]. The HOG vectors were extracted at a single scale to minimize computational complexity, with default parameters as listed in appendix A. The proposed IQA algorithm is termed the HOG Mean Squared Error (HMSE) and is defined as follows:

consider two discrete HOG vectors $x = \{x_i | i = 1, 2, \ldots, N\}$ and $y = \{y_i | i = 1, 2, \ldots, N\}$, where N is the number of entries in each vector, and x_i and y_i are the gradient magnitudes of the i^{th} entry in vectors x and y respectively. Then HMSE is given by:

$$HMSE = \frac{1}{N} \sum_{i=1}^{N} (x_i - y_i)^2 \tag{3.1}$$

3.5 Experimental Results

3.5.1 Database Creation

In order to evaluate the proposed metric against existing IQA algorithms a database of distorted images was created. The approach used was to consider the effects of 3 different types of degradation on commonly used pedestrian detection algorithms. There are a number of pedestrian detection databases freely available for use by the research community, for example [113] and [114]. In this work, both the Penn-Fudan database for pedestrian detection and segmentation [115] and the set of test images provided in the Institut National de Recherche en Informatique et en Automatique (INRIA) database [85] were used as the reference data sets. The Penn-Fudan database consists of scenes taken around a college campus and on urban streets. The primary reference (undistorted data) set consists of 173 images, each of which contains at least one pedestrian. The entire set of positive test images from the INRIA data set further complement the undistorted data set. The INRIA test data set consists of 288 images consisting of many scenes with multiple pedestrians and includes challenging examples of pedestrians partly hidden from view either by other pedestrians or vehicles. Furthermore, the data set includes a wide range of environmental lighting conditions including pedestrians standing in dark shadow or bright sunlight. The combined reference set thus contains a total of 461 images in which there are approximately 1,000 annotated pedestrians, including a variety of conditions, however, the reference set does not include images taken in rain, fog or any other extreme weather conditions. Such data sets are not readily available. In any case, the purpose of this experiment was to determine the effect of degradations introduced by the image acquisition and in-vehicle transmission system rather than environmental distortions on the performance of pedestrian detection algorithms, i.e. any artifact that distorts the displayed image from that of the ideal or “natural image”, that, for example, the driver can see through his or her window.

The types of image degradation that are most prevalent in automotive images are compression artifacts and noise [11, 13]. For each of the three degradations in the experiment, a model with only one variable parameter was used in order to simplify control of the degradation level. Forty values for the variable parameter in each model were chosen to ensure that the data set of degraded images contained a wide distribution of quality levels with reasonable granularity. A similar approach has been utilized in a number of image quality databases including [116] and [117]

Motion JPEG is currently one of the most popular methods of compression for real-time automotive video [118], however JPEG compressed images are known to exhibit blocking artifacts [119]. Blocking artifacts are common to all block-DCT based image compression techniques. In JPEG compression the discrete cosine transform is typically performed on 8×8 pixel blocks in each image frame and the coefficients in each block are quantized separately [120]. This is done by simply dividing each component in the frequency domain by a constant for that component, and then rounding to the nearest integer. The JPEG standard [121] specifies the 8x8 quantization matrix used at the quantization stage. The rounding operation is the main lossy operation in the compression process, assuming the DCT computation is performed with sufficiently high precision. As a result, it is typically the case that many of the higher frequency components are rounded to zero, while lower frequency components tend to become small positive or negative numbers, which take fewer bits to represent. The quantization process, however, can result in artificial horizontal and vertical borders between each block. Blocking artifacts can also be caused by transmission errors, which can affect entire regions of blocks in an image. Forty sets of JPEG compressed images were created by applying different CRs to the reference set. Adapting the level of compression in the model was achieved by weighting the quantization matrix by a quality factor Q, where Q = 100 represents the lowest rate of compression (and, hence highest visual quality), and Q = 0 represents the highest rate of compression (and hence lowest visual quality).

Although JPEG compression is the current industry standard, a more recent compression algorithm is the JPEG2000 (JP2) standard. It is therefore of interest to compare the performance of both compression algorithms. An image distortion found more prevalently in JP2 compression is ringing [72]. Ringing is caused by the quantization of high frequency coefficients in transform coding and is characterized by ripples around sharp edges. An example of this phenomenon was shown in Figure 1.2. Forty CRs were applied to the reference set to create sets of JP2 compressed images.

Sensor noise is a further distortion that can affect automotive image quality. Forty sets of noise corrupted images were created by applying AWGN to the reference images; an example can be seen in Figure 1.1. The variance of the noise was altered in order to create different noise levels, with higher variance corresponding to higher levels of AWGN. Overall, this process produced 120 sets of the reference data set consisting of varying levels of the three distortion types. An identical distortion parameter was applied to all images in any particular set. The complete data set therefore consists of over 55,000 images in which there are a total of over 120,000 annotated pedestrians under varying levels of distortion. The specific quality parameters used for each distortion type are detailed in appendix A. Matlab scripts for generating these images are included in appendix B.

3.5.2 Evaluating Detector Performance

The performance of three pedestrian detection algorithms, namely HOG+SVM [85], DPM [99] and ICF [86] were evaluated on the data set of degraded images. The default parameters for each algorithm were used and each detector was trained on the INRIA training set. Details of the default parameters for each algorithm are described in appendix A. The method described by the Pattern Analysis, Statistical Modelling and Computational Learning (PASCAL) visual objects classes challenge [111] was followed to evaluate correct detection. A detection is considered to be correct if the area of overlap a between the predicted bounding box B_p and ground truth bounding box B_t exceeds 0.5, as calculated by the formula:

$$a = \frac{area(B_p \cap B_t)}{area(B_p \cup B_t)} \tag{3.2}$$

The precision recall curve was then computed for the reference set and all sets of degraded images, where recall is defined as the number of true positives divided by the total number of pedestrians, and precision is the number of true positives divided by the total number of classifications returned by the algorithm (i.e. the number of true positives added to the number of false positives).

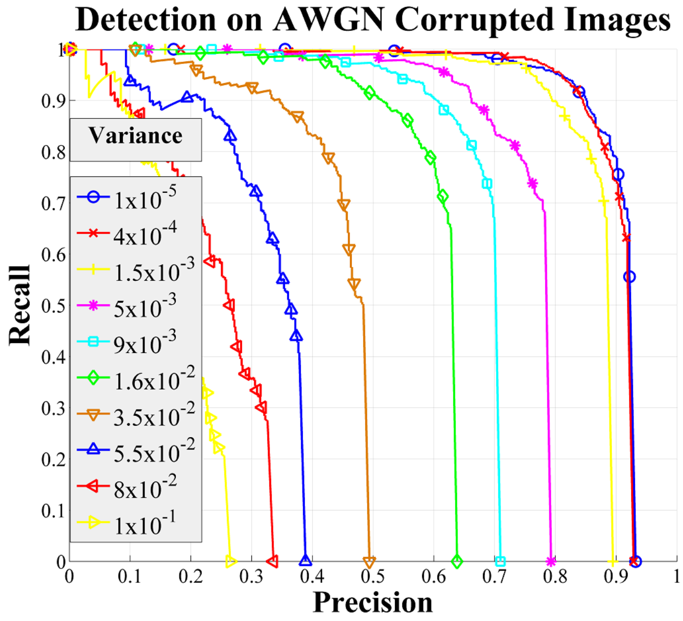
The Average Precision (AP) [122] is a useful metric for performance evaluation since it represents algorithm performance by a single value which reflects the shape of the precision-recall curve. AP is calculated by ranking the test set by classifier score, computing the precision at each rank and then averaging the result. In Figures 3.5-3.7 the ICF detection algorithm's performance for each distortion type is displayed. A higher area under the precision-recall curves represents higher detection performance. Figure 3.5(a) illustrates the precision recall curves for various levels of AWGN degraded reference sets. Figures 3.6(a) and 3.7(a) show the precision-recall curves for different CRs of JPEG and JP2 compression respectively, where CR is defined as the ratio between the uncompressed and compressed image sizes. Very similar results were found for both HOG+SVM and DPM detection algorithms. Pearson's Linear Correlation Coefficients (PLCCs) of algorithm performance (measured in AP) against degradation levels for all three algorithms are displayed in Table 3.2.

Interestingly, different distortions have very different effects on the performance of the pedestrian detection algorithm. The presence of AWGN in an image severely

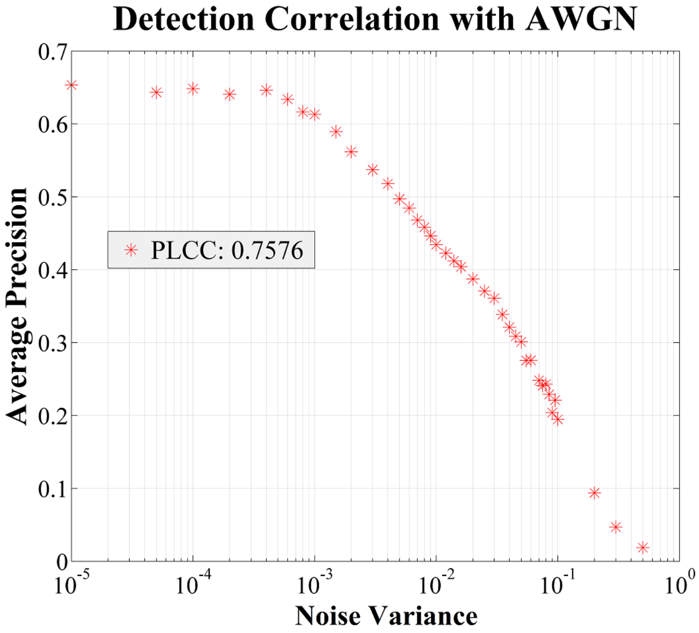
Table 3.2: PLCC of Algorithm Performance Against Degradation

	AWGN	**JPEG**	**JP2**
HOG+SVM	0.8216	0.6309	0.2314
ICF	0.7576	0.5410	0.1922
DPM	0.7913	0.6295	0.1791

impacts the recall of the detection algorithm. From Figure 3.5(b) it can be observed that a significant reduction in AP occurs between variance values of 10^{-3} and 10^{-2}. These values correspond to PSNR scores of approximately 30 and 27dBs respectively. Furthermore, for a large range of variances, the AP decreases linearly with increased AWGN. This result is evident from the high correlation between AWGN corruption and detection performance for all detection algorithms, however the same is not true of either JPEG or JP2 compression, both of which have well-defined knee points in the AP-CR curves. It can be seen from Figures 3.6(b) and 3.7(b) that high levels of compression can be achieved with little reduction in detection performance. It is also evident that blocking artifacts introduced by JPEG compression have a more serious impact on pedestrian classification than the ringing artifacts introduced by JP2 compression, affecting algorithm performance at much lower levels of compression.

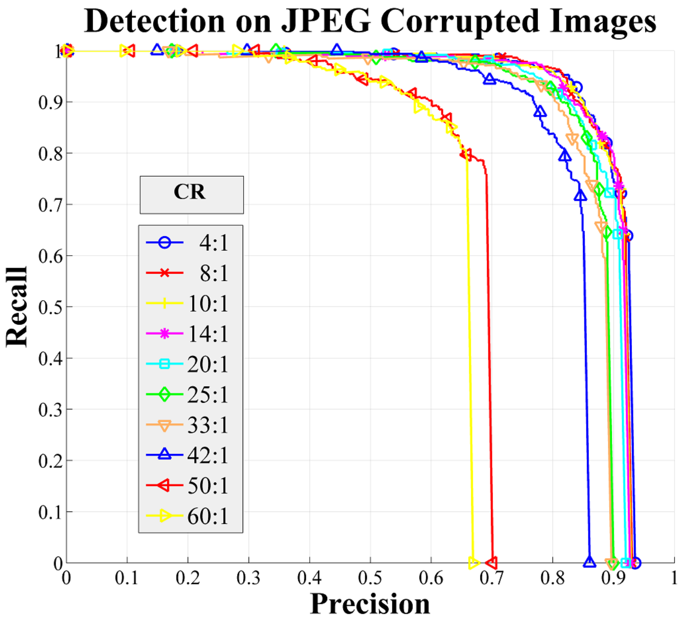


(a)

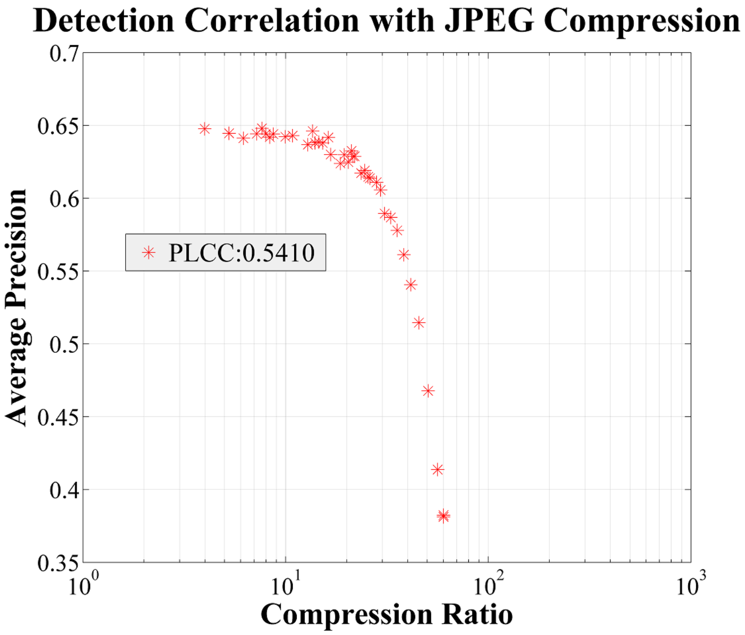


(b)

Figure 3.5: The precision recall curves for AWGN at different noise levels are shown in (a). In (b) the relationship between noise variance and average precision is shown.

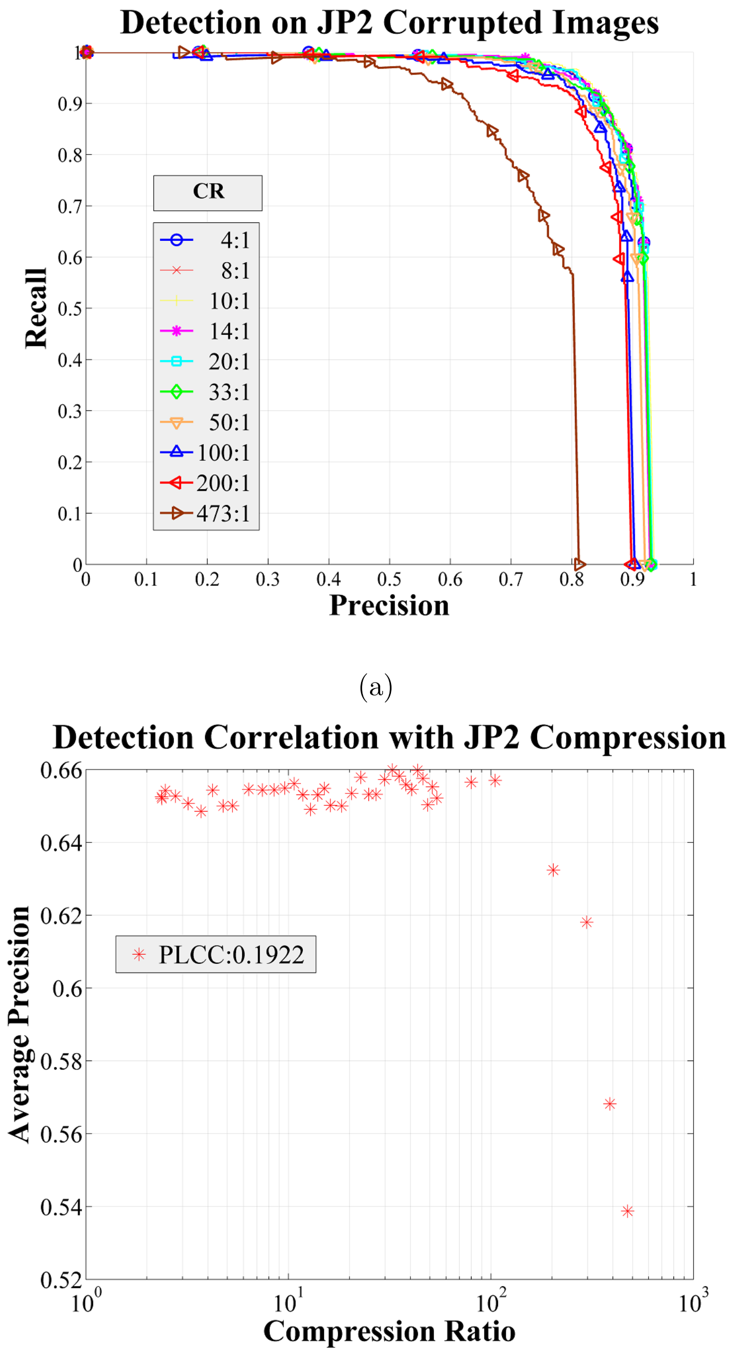


(a)



(b)

Figure 3.6: The precision recall curves for levels of JPEG compression at different CRs are shown in (a). In (b) the correlation between compression and average precision is shown.



(a)

(b)

Figure 3.7: The precision recall curves for levels of JP2 compression are shown in (a). In (b) the correlation between compression and average precision is shown.

3.5.3 Influence of Algorithm Retraining

Apart from choice of features, algorithm training is a significant factor affecting overall detection performance [123]. The influence of classifier retraining on the performance of all three detection algorithms was investigated. Each algorithm was retrained multiple times on varying quality versions of the INRIA training set [85] in order to assess the effect of training on overall detection performance. Details of the training sets used are contained in appendix A. Only the results for the ICF algorithm are presented and analysed in detail here as the behaviour for the other two detection algorithms was quite similar.

To compare detectors, miss rates were plotted against False Positives Per Image (FPPI) using log-log plots. This method was chosen instead of precision-recall curves since there is typically an upper limit on the acceptable FPPI rate for pedestrian detection in automotive applications. Detector performance is summarized using the Log-Average Miss Rate (LAMR), as used in [87]. The LAMR is computed by averaging the miss rates at 9 evenly spaced points (on a logarithmic scale) from 10^{-2} to 10^{0} FPPI. Conceptually, the LAMR is similar to the average precision reported for the PASCAL challenge [111], since it represents the entire curve by a single reference value. As curves are somewhat linear in this range (e.g., see Fig. 3.8), the LAMR is similar to the miss rate at 10^{-1} FPPI, but in general gives a more stable and informative assessment of performance.

The results for the ICF detection algorithm are shown in Figures 3.8 - 3.12 where lower curves represent lower FPPI and hence better detection performance. Training the algorithm on the undistorted INRIA training data set provides optimal perfor-

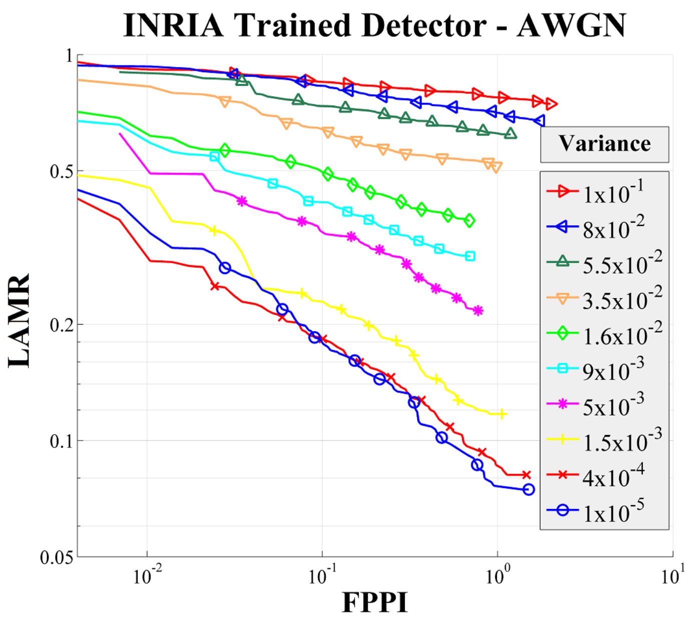
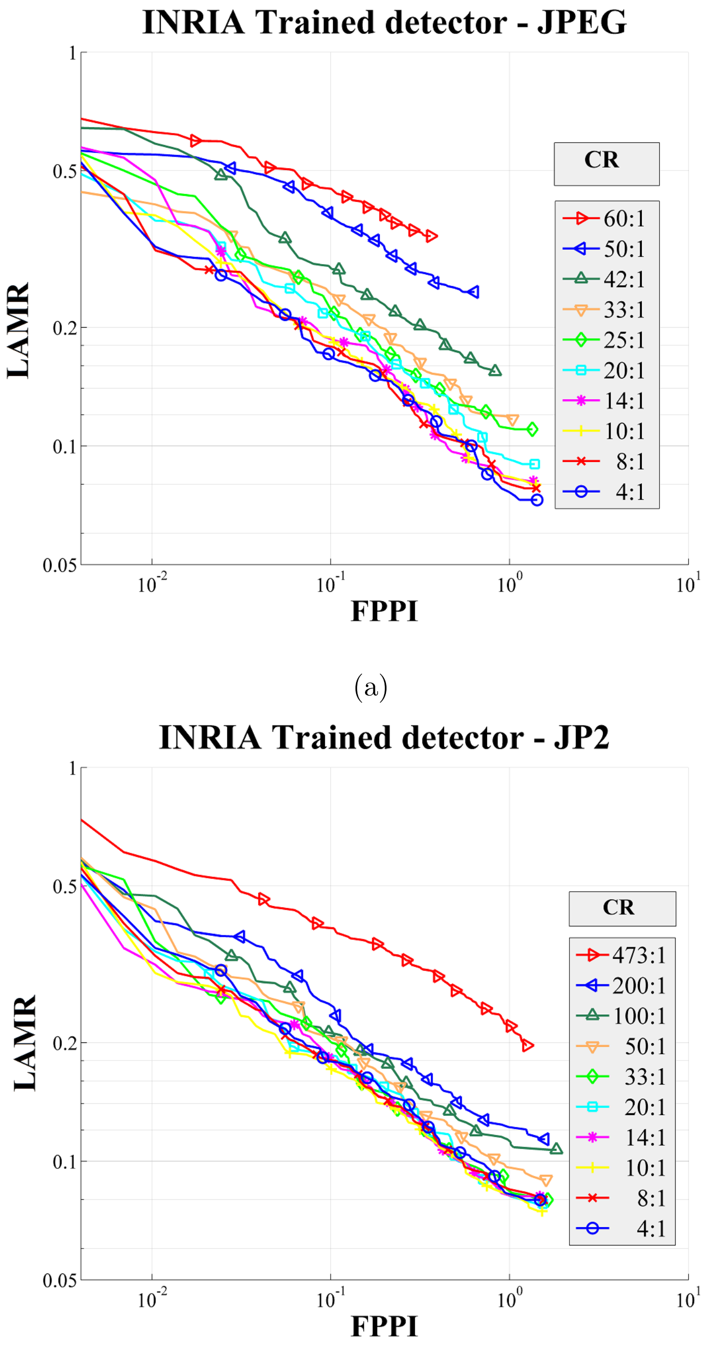


Figure 3.8: INRIA trained detector performance on varying levels of AWGN degradation.

mance on the reference set with a LAMR of 15.73%, however this classifier is not robust to image degradations such as AWGN or compression artifacts, as shown in Figures 3.8 and 3.9.

Notice that the LAMRs increase significantly when degradations are added to the images. On the other hand, training the classifier exclusively with poor quality samples leads to the greatest reduction in miss rate for poor quality images.

Figure 3.10(a) illustrates the detection performance of the ICF detection algorithm trained on low quality AWGN samples, on the AWGN data set. Not surprisingly, the detection performance of this classifier on poor quality AWGN images is superior to the reference-trained classifier, reducing LAMRs by over 30% in the case of images corrupted with the highest levels of AWGN (variance 1×10^{-1}).



(a)

(b)

Figure 3.9: INRIA trained ICF detector performance on JPEG (a) and JP2 (b) compressed images.

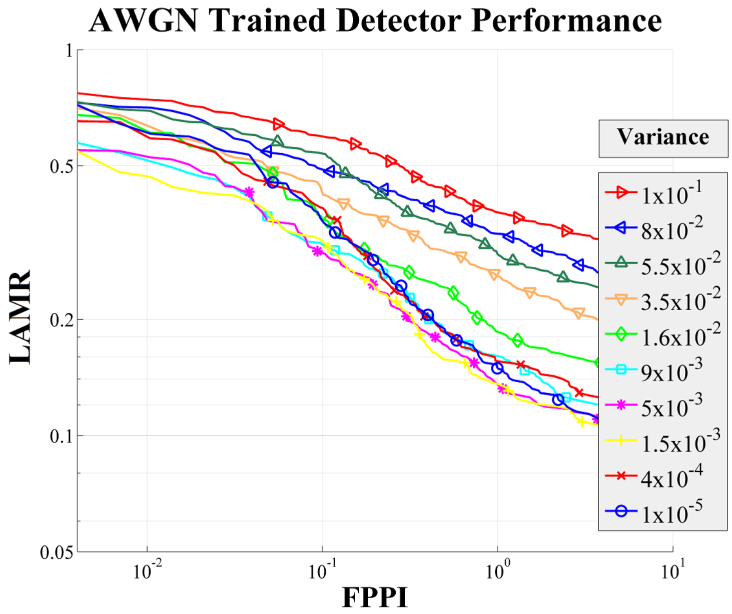


Figure 3.10: AWGN trained ICF detector performance on varying levels of AWGN degradation.

Improvements on heavily compressed images can also be achieved by training with similarly compressed images. For example, in Figure 3.11 the LAMR on the highest level of JPEG compressed images (CR:60:1) is reduced by over 10% compared with the reference-trained classifier.

Unfortunately, these improvements come at a high cost to detection performance on higher quality images, increasing LAMRs on the reference set by 16% and 23% for AWGN and JPEG trained classifiers respectively. In Figure 3.12 the detection results of classifiers trained using different paradigms, and tested on the entire data set are shown. The experiments reveal that a wide range of image quality in the training set optimizes algorithm performance across the entire test data set. Classifiers trained with multiple levels of quality outperform the reference, AWGN and compression-

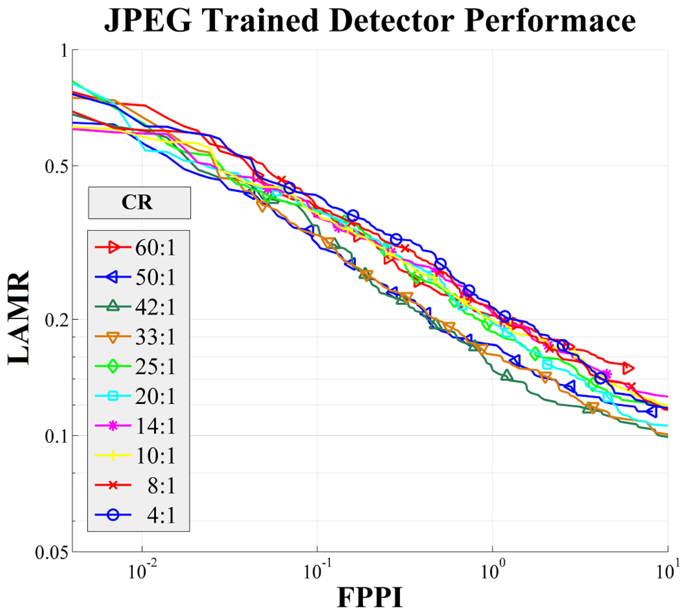


Figure 3.11: JPEG trained ICF detector performance on varying levels of JPEG compression.

trained classifiers, reducing LAMRs on the entire database by over 3%. Similar performance improvements were found for both the HOG+SVM and DPM algorithms. For the remainder of the experiments in this chapter, detection classifiers trained with multiple levels of quality are used, since they offer the best detection performance on the entire data set.

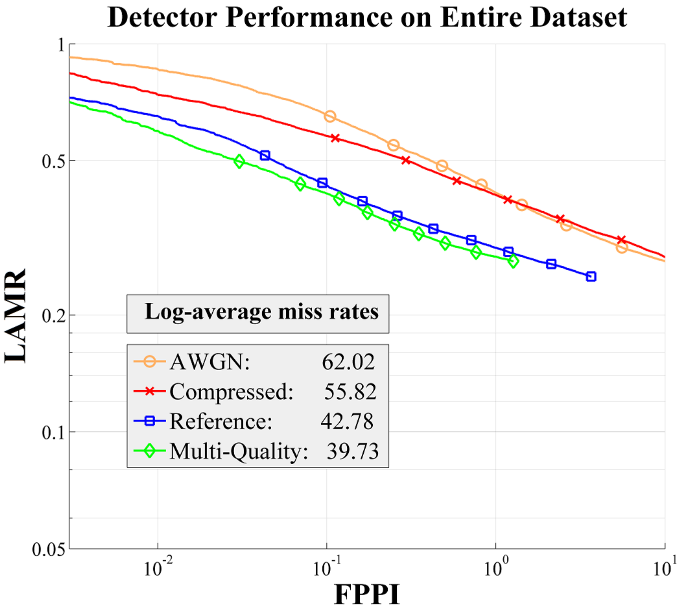


Figure 3.12: Performance of classifiers trained in different ways. Training the ICF detector with different levels of image quality influences detection performance. Optimal performance was achieved on the data set by including a wide variety of degradations in the training set.

3.5.4 Proposed IQA Algorithm Performance

The performance of the proposed HMSE metric was evaluated against eight existing IQA metrics described in Chapter 2. For each metric and each distortion type, the relevant images were sorted into bins based on their metric score. For example, since SSIM returns a score between zero and one, images were separated into 100 bins of width 0.01. In practice, not all bins contained images since an SSIM score of below 0.2 was not achievable even in the presence of extreme distortion; however the size of the database ensured that at least 60 consecutive bins contained a minimum of 100 images in all of the tests, in order to maximise coverage. In order to evaluate IQA algorithm performance, the AP and average IQA score for each bin was computed. Optimally trained HOG+SVM, DPM and ICF classifiers were used to assess detection performance. PLCC was then computed between IQA scores and detection performance for each algorithm in order to determine the strength of each IQA algorithm as a predictor of detection performance.

Table 3.3: PLCC of HOG+SVM Detection Performance Against Metric Scores

	PSNR	**SSIM**	**VSNR**	**VIF**	**NQM**	**SR-SIM**	**IWSSIM**	**FSIM**	**HMSE**
JPEG	0.8107	0.8491	0.8214	0.8301	0.8255	0.8796	0.8817	0.9017	0.9343
JP2	0.5321	0.5487	0.5614	0.5118	0.6501	0.6372	0.5518	0.6054	0.8213
AWGN	0.9218	0.9497	0.9472	0.9443	0.9681	0.9600	0.9732	0.9828	0.9806
ALL	0.8721	0.8973	0.8802	0.8514	0.9187	0.9302	0.9217	0.9489	0.9659

Table 3.4: PLCC of DPM Detection Performance Against Metric Scores

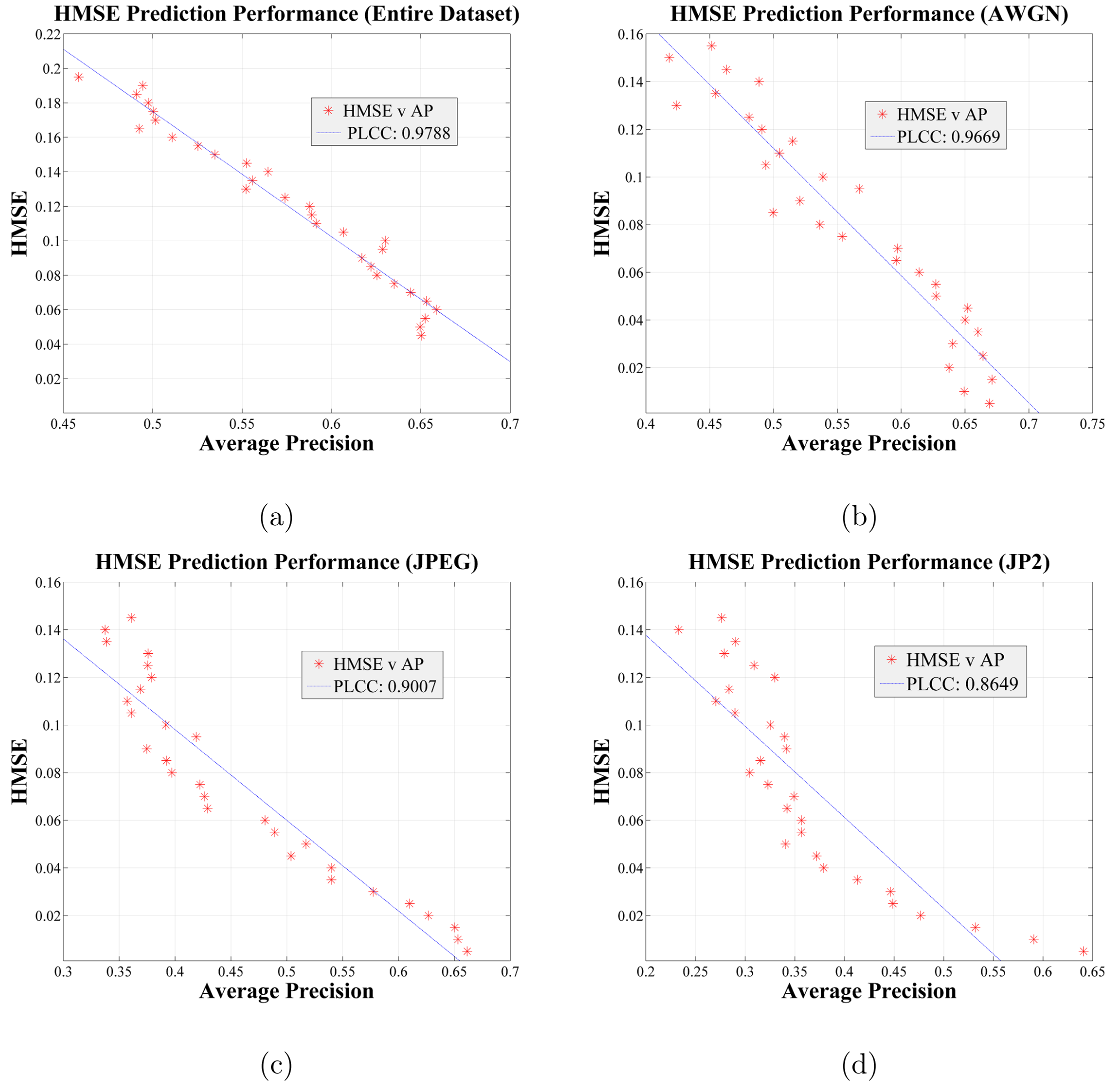
	PSNR	**SSIM**	**VSNR**	**VIF**	**NQM**	**SR-SIM**	**IWSSIM**	**FSIM**	**HMSE**
JPEG	0.8018	0.8517	0.8661	0.9016	0.9238	0.9234	0.9185	0.9012	0.9436
JP2	0.4011	0.4784	0.4318	0.4819	0.6318	0.7063	0.5527	0.5233	0.8501
AWGN	0.8710	0.9220	0.8911	0.9155	0.9737	0.9584	0.9673	0.9611	0.9617
ALL	0.8344	0.8986	0.8761	0.8902	0.9317	0.9440	0.9138	0.9254	0.9632

Table 3.5: PLCC of ICF Detection Performance Against Metric Scores

	PSNR	**SSIM**	**VSNR**	**VIF**	**NQM**	**SR-SIM**	**IWSSIM**	**FSIM**	**HMSE**
JPEG	0.6348	0.7571	0.7071	0.7272	0.8316	0.8338	0.8417	0.8222	0.9007
JP2	0.4329	0.4704	0.4881	0.4911	0.5675	0.7536	0.5017	0.5265	0.8649
AWGN	0.8316	0.9350	0.8851	0.8598	0.9754	0.9650	0.9812	0.9767	0.9692
ALL	0.6912	0.9042	0.7810	0.7581	0.9112	0.9442	0.9628	0.9601	0.9788

The results, shown in Figure 3.13, and summarized in Tables 3.3 to 3.5 illustrate the strong correlation between HMSE and pedestrian detection performance across all types of image degradations in the data set. Over the entire data set, HMSE achieves a PLCC of over 0.96 for all pedestrian detection algorithms. On the subsets of the data set corresponding to different distortions, the HMSE metric correlates best with AWGN distortion, again achieving PLCCs of over 0.96 for all detection algorithms. In general, AWGN is the "easiest" distortion to predict, with all IQA algorithms achieving high correlation with detection performance. Predicting algorithm performance on JPEG and JP2 compressed images is more difficult, particularly for JP2 compression. Most of the IQA algorithms achieved only weak correlation with detection performance in this case. This is largely due to the fact that ringing artifacts, which can cause false positives in detection, often occur at frequencies that are difficult for a human viewer to perceive. Many common IQA algorithms may deliberately disregard such distortions as part of their HVS model.

Despite these challenges, some of the recent IQA algorithms, particularly SR-SIM correlate well with detection performance. This is likely because SR-SIM determines quality by measuring the difference between the reference and degraded images' saliency maps. In SR-SIM the saliency maps are derived from gradient images extracted from the test images. Changes to low-level features will hence change the derived saliency maps and increase the "distance" between reference and degraded images.



(a)

(b)

(c)

(d)

Figure 3.13: The correlation of HMSE with ICF detection performance for the entire data set (a), the subset of AWGN corrupted images (b), and for JPEG (c), and JP2 (d) compressed subsets of the data set.

3.5.5 Statistical Significance

In Tables 3.3 to 3.5 the correlation coefficients of HMSE and eight other IQA algorithms were presented. Although the HMSE metric generally has higher correlation with algorithm performance than other metrics, this deserves further consideration. In this section, the statistical significance of these differences in correlation is examined. Given two correlation coefficients and their associated sample sizes, the Fisher test [124] determines whether the two coefficients are statistically different from each other. The correlation coefficients computed for each detection algorithm were analyzed for significance at the 95% confidence level. The results are tabulated in Tables 3.6 to 3.8 in a presentation style similar to [77]. Each row/column entry consists of four symbols, which represent the entire, AWGN, JPEG, and JP2 data sets, respectively. A one indicates the row IQA algorithm is statistically superior to the column IQA algorithm. A dash (-) indicates statistical equivalence, while a zero indicates that the row algorithm is statistically inferior to the column algorithm. The concentration of "1"s in the first row (for HMSE) indicates that the prediction performance of HMSE is statistically better than many of the other algorithms tested. For example, on the entire data set, HMSE is statistically better than all other algorithms for ICF performance prediction, and statistically better than seven of eight algorithms for both DPM and HOG+SVM performance prediction.

Table 3.6: HOG+SVM: Statistical Significance Matrix

All/AWGN/JPEG/JP2	HMSE	IWSSIM	FSIM	SR-SIM	SSIM	NQM	VIF	VSNR	PSNR
HMSE	-	1/-/1/1	-/-/-/1	1/1/1/1	1/1/1/1	1/-/1/1	1/1/1/1	1/1/1/1	1/1/1/1
IWSSIM	0/-/0/0	-	-/0/-/-	-/-/-/-	-/1/-/-	-/-/-/-	1/1/-/-	-/1/-/-	-/1/-/-
FSIM	-/-/-/0	-/-/-/-	-	-/1/-/-	1/1/-/-	-/-/1/-	1/1/1/-	-11/1/-	1/1/1/-
SR-SIM	0/0/0/0	-/-/-/-	-/0/-/-	-	-/-/-/-	-/-/-/-	1/-/-/-	1/-/-/-	1/1/-/-
SSIM	0/0/0/0	-/0/-/-	0/0/-/-	-/-/-/-	-	-/-/-/-	-/-/-/-	-/-/-/-	-/-/-/-
NQM	0/-/0/0	-/-/-/-	-/-/0/-	-/-/-/-	-/-/-/-	-	1/1/-/-	-/-/-/-	-/1/-/-
VIF	0/0/0/0	0/0/-/-	0/0/0/-	0/-/-/-	-/-/-/-	0/0/-/-	-	-/-/-/-	-/-/-/-
VSNR	0/0/0/0	-/0/-/-	0/0/0/-	0/-/-/-	-/-/-/-	-/-/-/-	-/-/-/-	-	-/-/-/-
PSNR	0/0/0/0	-/0/-/-	0/0/0/-	0/0/-/-	-/-/-/-	-/0/-/-	-/-/-/-	-/-/-/-	-

Table 3.7: DPM: Statistical Significance Matrix

All/AWGN/JPEG/JP2	HMSE	IWSSIM	FSIM	SR-SIM	SSIM	NQM	VIF	VSNR	PSNR
HMSE	-	1/-/-/1	1/-/1/1	-/-/-/1	1/1/1/1	1/-/-/1	1/1/1/1	1/1/1/1	1/1/1/1
IWSSIM	0/-/-/0	-	-/-/-/-	-/-/-/-	-/1/1/-	-/-/-/-	-/1/-/-	-/1/-/-	1/1/1/-
FSIM	0/-/0/0	-/-/-/-	-	-/-/-/0	-/1/-/-	-/-/-/-	-/1/-/-	-/1/-/-	1/1/1/-
SR-SIM	-/-/-/0	-/-/-/-	-/-/-/1	-	1/1/1/1	-/-/-/-	1/1/-/-	1/1/1/1	1/1/1/1
SSIM	0/0/0/0	-/0/0/-	-/0/-/-	0/0/0/0	-	-/0/0/-	-/-/-/1	-/-/1/-	-/-/-/-
NQM	0/-/-/0	-/-/-/-	-/-/-/-	-/-/-/-	-/1/1/-	-	-/1/-/-	1/1/1/1	1/1/1/1
VIF	0/0/0/0	-/0/-/-	-/0/-/-	0/0/-/0	-/-/-/-	-/0/-/-	-	-/-/-/-	-/-/1/-
VSNR	0/0/0/0	-/0/-/-	-/0/-/-	0/0/0/0	-/-/-/-	0/0/0/0	-/-/-/-	-	-/-/-/-
PSNR	0/0/0/0	0/0/0/-	0/0/0/-	0/0/0/0	-/-/-/-	0/0/0/0	-/-/0/-	-/-/-/-	-

Table 3.8: ICF: Statistical Significance Matrix

All/AWGN/JPEG/JP2	HMSE	IWSSIM	FSIM	SR-SIM	SSIM	NQM	VIF	VSNR	PSNR
HMSE	-	1/-/-/1	1/-/1/1	1/-/-/1	1/1/1/1	1/-/1/1	1/1/1/1	1/1/1/1	1/1/1/1
IWSSIM	0/-/-/0	-	-/-/-/-	-/1/-/0	1/1/-/-	1/-/-/-	1/1/1/-	1/1/1/-	1/1/1/-
FSIM	0/-/0/0	-/-/-/-	-	-/-/-/0	1/1/-/-	1/-/-/-	1/1/-/-	1/1/1/1-	1/1/1/-
SR-SIM	0/-/-/0	-/0/-/1	-/-/-/1	-	1/1/-/1	1/-/-/1	1/1/-/1	1/1/1/1	1/1/1/1
SSIM	0/0/0/0	0/0/-/-	0/0/-/-/	0/0/-/0	-	-/0/-/-	1/1/-/-	1/1/-/-	1/1/-/-
NQM	0/-/0/0	0/-/-/-	0/-/-/-	-/-/-/0	-/1/-/-	-	1/1/-/-	1/1/1/-	1/1/1/-
VIF	0/0/0/0	0/0/0/-	0/0/-/-	0/0/-/0	0/0/-/-	0/0/-/-	-	-/-/-/-	-/-/-/-
VSNR	0/0/0/0	0/0/0/-	0/0/0/-	0/0/0/0	0/0/-/-	0/0/0/-	-/-/-/-	-	-/-/-/-
PSNR	0/0/0/0	0/0/0/-	0/0/0/-	0/0/0/0	0/0/-/-	0/0/0/-	-/-/-/-	-/-/-/-	-

3.5.6 Computational Complexity

In Table 3.9 the computational complexity of each of the IQA algorithms is evaluated. The time to compute the quality score of a typical test image of resolution 375 × 512 pixels on a 3.3GHz PC with 8GB of RAM operating on a single core is reported. Unoptimized Matlab® code was used for all algorithms to ensure a fair comparison. Notwithstanding the use of unoptimized code, it can be seen that HMSE has low computational complexity, ranking second of all algorithms tested with respect to running speed. Only PSNR offers lower computational complexity, however HMSE achieves statistically better performance prediction than PSNR for all detection algorithms in this experiment. Furthermore, the computation time reported for HMSE includes extraction of the HOG features. In current automotive vision systems, the HOG feature may already be computed for detection algorithms. HMSE could potentially be incorporated into such systems with little additional computational cost.

Table 3.9: Complexity Analysis of HMSE

Algorithm	**Time(seconds)**
PSNR	0.01
HMSE	0.03
SR-SIM	0.06
SSIM	0.12
MSSIM	0.18
NQM	0.46
VSNR	0.50
IWSSIM	0.67
IFC	1.14
VIF	1.15

3.6 Discussion and Conclusions

A study on the influence of image artifacts on pedestrian detection performance was carried out in this chapter. A particularly interesting result is that different image artifacts have very different effects on the performance of image processing algorithms. This is due to the fact that different types of degradation influence low-level image features in characteristic ways, which in turn can impact the performance of higher level machine vision algorithms such as pedestrian detection that make use of particular features. Degradation in detection performance responded linearly with increased AWGN, however for compression, knee points in algorithm performance were found for both JPEG and JP2 compression algorithms; detection performance was maintained at a reasonably good level until CR exceeded a certain value. These results have applications in the system design stage as they could be used to optimize image capture parameters for machine vision performance. The results also highlight the impact of different compression algorithms on detection performance. For all three pedestrian detection algorithms tested, it was found that JPEG compression had a larger impact on detection performance than JP2 compression for similar CRs. This is perhaps not a surprising result since JP2 compression is a more sophisticated and recent algorithm than JPEG compression. Nevertheless, the difference between the algorithms is significant, with JP2 facilitating compression ratios an order of magnitude higher than JPEG without significantly impacting detection performance. It has been demonstrated that preservation of low level image features is critical in order to ensure robust performance of pedestrian detection algorithms.

Finally, a new IQA algorithm for performance prediction of pedestrian detection algorithms has been proposed. The proposed metric, HMSE, utilizes low level image gradients to capture the structural information inherent in a scene. Although relatively simple and computationally inexpensive, the metric correlates closely with algorithm performance across a range of different image degradations which are typically found in automotive vision systems. For the particular task of performance prediction of pedestrian detection algorithms, the metric statistically outperforms current state-of-the-art perceptual quality metrics such as SSIM, IWSSIM and FSIM.

Like many other algorithms, a disadvantage of HMSE is the need for a reference image, which limits the applications of the proposed quality metric to system design and experimental performance evaluation. In the next chapter NR approaches to pedestrian detection performance evaluation are discussed and a framework for optimizing detection classification on degraded video sequences is proposed.

Chapter 4

Performance Optimization using Natural Scene Statistics

4.1 Introduction

The previous chapter examined the effect of different degradations on pedestrian detection performance, and established some basic results relating to the relationship between image quality and detection performance. A new full-reference IQA metric for predicting pedestrian detection performance was also proposed. In this chapter, a much larger data set of automotive images and video sequences is created. It is further verified that different transmission induced quality degradations affect the performance of a pedestrian detection classifier, based on Integral Channel Features [86, 125], in characteristic ways. Furthermore, the effect of classifier retraining on improving the robustness of the detection algorithm to image quality degradations is examined in more detail. Then, a novel detection framework which utilizes NR image quality statistics, based on [81], to categorise distorted frames and enhance the performance of the detection algorithm through the use of multiple “distortion specific” classifiers is introduced. NR image quality assessment and distortion classification are combined with a multi-classifier approach to pedestrian detection in order to increase detection performance on degraded images. To ensure robust detection performance, the same setup as used for the best results in the original ICF paper [86], in which learning is achieved via discrete Adaboost, is used throughout this chapter. The outcome of this work opens up the possibility for development of a higher performing pedestrian detection system, suitable for real-time detection, where NR image quality is used to categorise the level of distortion and an appropriate classifier to maximise performance for distortion level is chosen accordingly.

4.2 Multi-classifier Detection Framework

4.2.1 Database Creation

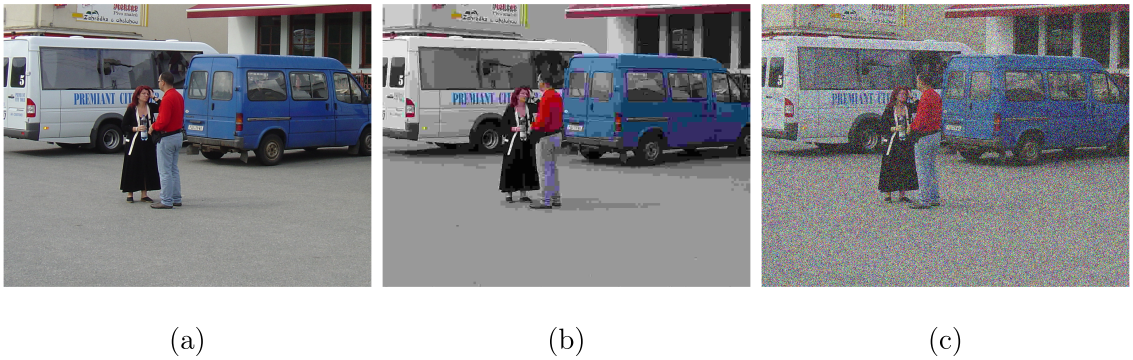
The majority of the experiments were performed on the INRIA [85] and Caltech [87, 126] data sets; however, in order to further demonstrate the performance of the proposed algorithm, it was also evaluated on real distortions using the pedestrian subset of the ChangeDetection (CD) 2014 data set [127].

The INRIA data set described in Chapter 3 consists of a total of 1805 pedestrians from a varied set of images. It is available at `http://pascal.inrialpes.fr/data/human/`, along with annotation files to download for research purposes. The more recent Caltech pedestrian detection benchmark consists of approximately 10 hours of

640 × 480 30 fps video taken from a vehicle driving through regular traffic in an urban environment. The Caltech data set is much larger than the INRIA data set as it consists of 350,000 labelled pedestrian bounding boxes in 250,000 frames. The pedestrians vary widely in appearance, pose and scale, thus the images collected are more representative of real world applications and allow for in-depth analysis of existing algorithms. The Caltech dataset and annotation files can be downloaded from `http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/`. Images are stored in SEQ files. A SEQ file is a series of concatenated image frames with a fixed size and header. It is is essentially the same as merging a directory of images into a single file. SEQ files are convenient for storing videos because no video codec is required, seeking is instant and exact, and SEQ files can be read on any operating system. The main drawback is that each frame is encoded independently, resulting in increased file size. Currently, either uncompressed, JPEG or PNG compressed frames are supported. The Caltech dataset is provided in uncompressed format and Matlab routines for reading/writing/manipulating SEQ files can be found in Piotr's Matlab Toolbox (version 3.20 or later) available here: `https://github.com/pdollar/toolbox`. The toolbox also includes routines for reading annotations.

The CD 2014 pedestrian detection data set contains 10 videos of pedestrians taken with different cameras ranging from low-resolution Internet Protocol (IP) cameras to mid-resolution camcorders. Due to the different lighting conditions and compression settings in each video, the CD videos have varying levels of compression artifacts and noise.

As in Chapter 3, video frames from both the INRIA and Caltech data sets were



(a) (b) (c)

Figure 4.1: Sample images from the test data set. A reference "RAW" image is shown in (a). "Blocking artifacts" caused by JPEG compression are evident on the side of the blue van in (b). AWGN, which can be introduced during image capture, also degrades the quality of the reference image, as illustrated in (c).

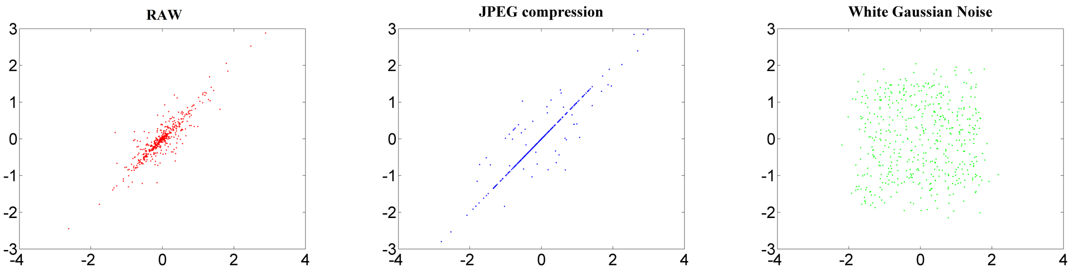
randomly separated into non-overlapping training (80%) and testing (20%) data sets before both JPEG compressed and AWGN degraded versions of each set were created. The original RAW versions of each data set were considered to be of ideal image quality, since they contained no apparent degradations introduced by either compression or noise artifacts. Twelve values were chosen for the variable parameter in each model to ensure that the new data set of degraded images contained a wide distribution of quality levels.

The same procedure as in Chapter 3 was used to add degradations to each image. Twelve batches of JPEG compressed images were created by sampling Q on a log scale as follows: Q = {60, 50, 30, 24, 19, 15, 12, 8, 6, 5, 4, 1}. The resulting batches had average compression ratios of {13, 15, 20, 23, 27, 31, 34, 42, 46, 51, 57 and 65} respectively. These values of Q were chosen to represent a wide range of subjective quality levels. Similar values have been chosen in other well-known image quality databases such as [116, 117, 16], however these databases use only 4-5 distortion levels. The database used in this research therefore provides a more diverse set of quality levels than in the current literature. The twelve levels of JPEG distortion are henceforth referred to as $JPEG_1$ to $JPEG_{12}$.

Both charge coupled device (CCD) and complementary metal oxide semiconductor (CMOS) technologies introduce sensor noise at image capture [128]. Typically, AWGN is used to model this image degradation. A database of noise corrupted images was created, wherein AWGN was added synthetically to the reference images. The variance of the noise was set to values from the following set: $\sigma^2 = \{5 \times 10^{-5}$, 0.0006, 0.0015, 0.003, 0.006, 0.009, 0.014, 0.025, 0.045, 0.085, 0.2, 0.3$\}$. Again, these values were chosen to represent a wide range of quality levels, and are consistent with the levels used in [116, 117, 16]. The twelve levels of AWGN are henceforth referred to as $AWGN_1$ to $AWGN_{12}$. Examples of each distortion are illustrated in Figure 4.1. These processes produced over 100,000 distorted images where a known distortion parameter was applied to each image. The complete data set therefore consists of over 200,000 annotated pedestrians with twelve levels for each type of image degradation.

4.2.2 Natural Scene Statistics

One of the established fundamental statistical properties of natural images is that two neighbouring pixels are correlated [74]. This can be easily demonstrated by a scatter plot of the grayscale values of neighbouring pixels sampled from RAW images from the database, which show a particular pattern as illustrated in Fig 4.2(a). Notice that distortions such as JPEG compression (4.2(b)) and AWGN (Fig. 4.2.c) alter the



(a) (b) (c)

Figure 4.2: Scatter plots of the grayscale values of neighbouring pixels are shown for RAW (a), JPEG compressed (b), and AWGN corrupted (c) images. The values have been scaled so that the mean is zero and variance one.

characteristic correlations of neighbouring pixels from that of their corresponding RAW images. These characteristic deviations in scene statistics provide us with information about the "naturalness" of the test image; such statistics derived from images are often referred to as natural scene statistics (NSS). Figure 4.3 shows a cross section of the correlation coefficients (with DC component removed) of a pixel with all neighbour pixels for a RAW, JPEG compressed and AWGN corrupted image from the data set. For RAW images, the coefficients are well modelled by a Laplacian distribution, however it is evident that each distortion affects the distribution in a characteristic way. For example, the addition of AWGN significantly decreases the correlation of neighbouring pixel values, while image compression removes some of the high frequency components in image blocks. The removal of these high frequencies tends to smooth each block, increasing a pixel's correlation with its neighbours. Such characteristic image statistics have been used for example in [76] to blindly categorise images by distortion type.

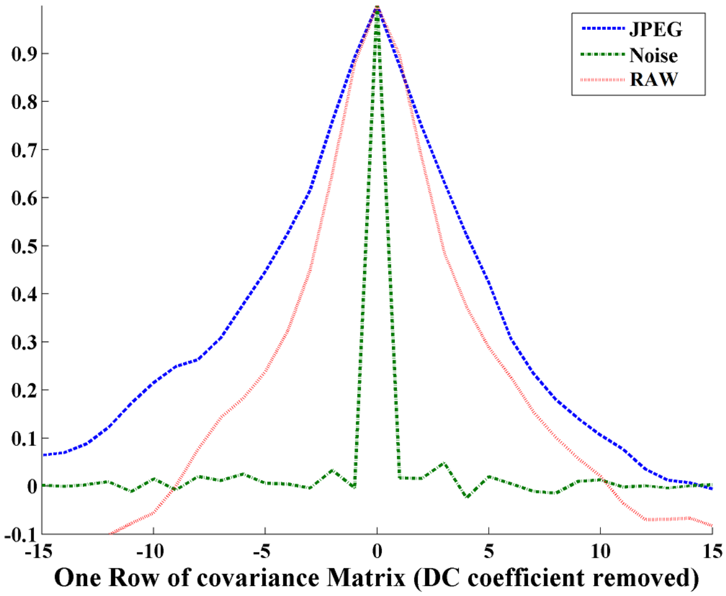


Figure 4.3: One row of the covariance matrices of neighbour pixels (with DC coefficient removed) of RAW, AWGN corrupted and JPEG compressed images from the data set.

The NIQE metric was used as a measure of perceived image quality since it is efficient to compute and has been shown to correlate highly with human opinion scores. The classical spatial NSS model adopted by NIQE is based on [129]. Each image (I) was pre-processed using local mean removal and normalization:

$$I(i,j) = \frac{(I(i,j) - \mu(i,j))}{(\sigma(i,j))} \tag{4.1}$$

where $i \in \{1, 2, \ldots, M\}$, $j \in \{1, 2, \ldots, N\}$ are spatial indices, M and N are the image dimensions, and μ and σ are the local mean and standard deviation respectively, weighted by a 2-D circularly symmetric Gaussian weighting function. The statistics of local image patches were then characterized by so-called "quality-aware" NSS features. Previous studies of NSS-based image quality have shown that the generalized Gaussian distribution effectively captures the behaviour of NSS features extracted from natural and distorted images. A simple model of the NSS features computed from natural image patches was obtained by fitting them to a multivariate Gaussian (MVG) model. The quality of a test image was then expressed as the distance (computed from equation 4.2) between an MVG fit of the NSS features extracted from the test image and an MVG model of the features extracted from a data set of natural images:

$$D(v_1, v_2, \Sigma_1, \Sigma_2) = \sqrt{((v_1 - v_2)^{\Gamma} \quad (\frac{(\Sigma_1 + \Sigma_2)}{2})^{-1} \quad (v_1 - v_2))} \tag{4.2}$$

where $v1$, $v2$ and Σ_1, Σ_2 are the mean vectors and covariance matrices of the natural

MVG model and the distorted image's MVG model respectively and Γ is the Gamma function:

$$\Gamma(a) = \int_0^{\infty} t^{a-1} e^{-t}\, dx \quad a > 0 \tag{4.3}$$

4.2.3 Allocating Images Into Quality Bins

In order to categorize the images in the data set, a model of the NSS features computed from the RAW images was obtained by fitting them to an MVG density according to [81]. One thousand of the INRIA training images were selected to generate the RAW MVG model. The NIQE score of every image in the training data set was then calculated. The images were binned according to their NIQE score and distortion type. In practice it was found that RAW images typically had an NIQE score below 4, JPEG compressed images typically ranged from 4 to 12, while AWGN corrupted images had a larger range of perceived image quality, typically ranging from 4 to over 16 on the NIQE scale (recall that the NIQE score for a particular test image is the distance between the RAW modelled MVG and the MVG computed from features extracted from the test image, therefore a lower value indicates better quality).

By binning the training images in the data set according to their distortion type and NIQE score the goal was to obtain new MVG models for specific distortion types and levels. After visually inspecting the distorted images and examining the impact of these distortions on pedestrian detection performance, bin widths of two NIQE units were chosen. A similar bin width was used in [89]. It was ensured that at least 1000 images were contained in each bin. The median score within each bin served as the representative quality score for that bin. MVG models were obtained from the samples in each bin, such that for any particular quality score, there existed corresponding MVG models of both noise and JPEG degraded images. In practice this means a total of 11 models were computed covering the entire spectrum of NIQE scores in the data set. These models were used to categorise distortion type. The bin levels for each distortion type are detailed in appendix A.

The quality scores of the degraded images were quantified by comparing the MVG model computed from the test images with that of the RAW model. All images below a particular quality threshold (NIQE = 4) were considered to have ideal quality and pedestrian detection was carried out with a classifier trained exclusively on non-degraded images. For images with quality scores above this threshold, the distance between the test image MVG and corresponding MVG models for each degradation was computed, in order to categorise the test image as either JPEG compressed or AWGN corrupted. Depending on the category of distortion, an appropriately trained classifier was then utilized for pedestrian detection, with the objective of improving performance on the test data set and increasing the robustness of the detection algorithm to image degradations. Similar NSS features have been used for distortion classification in [130] and [75] however these algorithms utilise SVMs for distortion classification. The approach is summarised in Fig. 4.4.

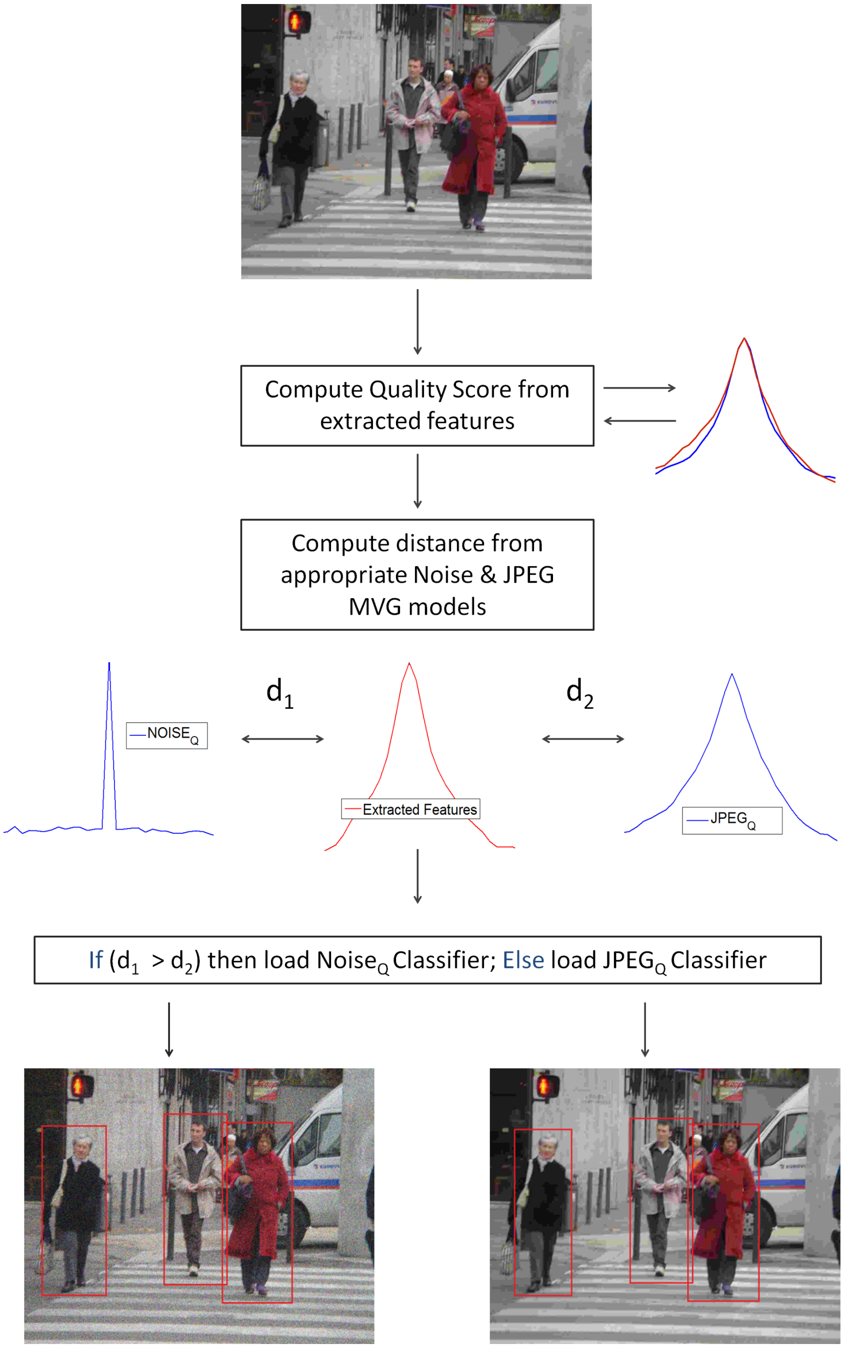

Compute Quality Score from extracted features
Compute distance from appropriate Noise & JPEG MVG models
NOISE$_Q$
d_1
Extracted Features
d_2
JPEG$_Q$
If (d_1 > d_2) then load Noise$_Q$ Classifier; Else load JPEG$_Q$ Classifier

Figure 4.4: A framework for improved pedestrian detection performance through NR image distortion categorization.

4.3 Experimental Results

In order for the method to be effective, accurate distortion classification is important, particularly for badly degraded images. We first used the INRIA and Caltech test sets described previously to evaluate the categorization technique. Categorizing high quality images is extremely difficult since very low levels of JPEG compression or AWGN distortion typically do not introduce perceptual image distortions and consequently do not significantly alter the scene statistics used for quality assessment. However, equally, such distortions have little effect on the performance of pedestrian detection algorithms. For example in Table 4.1, the LAMR of the detection algorithm is reported for each batch of INRIA distorted images (where lower LAMR represents higher detection performance). It is evident that the lowest two levels of each distortion introduce very little degradation in detection performance.

Table 4.1: Detection Performance on the distorted database (LAMR)

Distortion level	AWGN	JPEG
0 (undistorted)	18.17	18.17
1	17.64	18.00
2	18.91	18.26
3	22.15	19.62
4	28.59	22.28
5	37.64	20.73
6	42.65	23.02
7	47.90	22.10
8	57.62	26.73
9	70.83	29.90
10	83.95	30.29
11	95.00	37.82
12	97.60	45.34

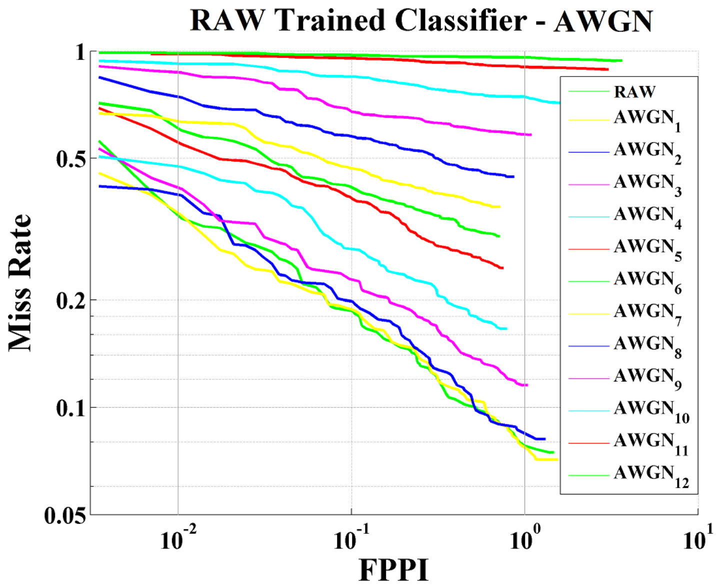
For this reason, images with an NIQE score below 4 were considered to have ideal quality, corresponding to 7% of the images in the distorted test set. The majority (76% of INRIA and 99% of Caltech data sets) of RAW classified images came from the first two levels of each distortion, representing the highest image qualities, while almost all RAW classified images came from the first four levels of each distortion type (99% INRIA, 100% Caltech). Optimal pedestrian detection performance on this subset of RAW classified images was achieved by using a classifier trained exclusively on the undistorted data set, hence the proposed detection framework remains appropriate for this subset of images. The classification results from the distorted image data set are reported in Table 4.2 with the correct classification type for each distortion level highlighted in bold font. The results of the remaining images in each data set are summarized below:

- Correctly classified distortions: 99.37% (INRIA), 99.74% (Caltech)
- JPEG incorrectly classified as AWGN: 0.006% (INRIA), 0% (Caltech)
- AWGN incorrectly classified as JPEG: 1.10% (INRIA), 0.28% (Caltech)

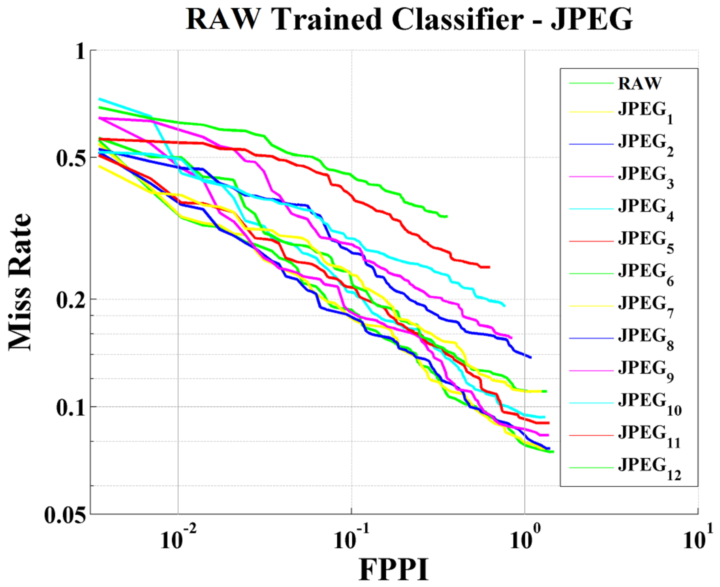
Table 4.2: Classification performance on INRIA and Caltech data sets

	Classification Percentages					
	RAW		AWGN		JPEG	
Quality\data set	INRIA	Caltech	INRIA	Caltech	INRIA	Caltech
Undistorted	**98.26**	**96.34**	0	0	1.74	3.66
$AWGN_1$	67.71	86.58	**19.1**	**10.81**	13.19	2.61
$AWGN_2$	11.46	1.64	**88.54**	**98.36**	0	0
$AWGN_3$	2.08	0	**97.92**	**100**	0	0
$AWGN_4$	0.35	0	**99.65**	**100**	0	0
$AWGN_5$	0	0	**100**	**100**	0	0
$AWGN_6$	0	0	**100**	**100**	0	0
$AWGN_7$	0	0	**100**	**100**	0	0
$AWGN_8$	0	0	**100**	**100**	0	0
$AWGN_9$	0	0	**100**	**100**	0	0
$AWGN_{10}$	0	0	**100**	**100**	0	0
$AWGN_{11}$	0	0	**100**	**100**	0	0
$AWGN_{12}$	0	0	**100**	**100**	0	0
$JPEG_1$	65.97	87.87	0.35	0	**33.68**	**12.13**
$JPEG_2$	56.6	2.01	0.35	0	**43.06**	**97.99**
$JPEG_3$	39.24	0.25	0	0	**60.76**	**99.75**
$JPEG_4$	20.14	0.55	0	0	**79.86**	**99.45**
$JPEG_5$	1.39	0	0	0	**98.61**	**100**
$JPEG_6$	0	0	0	0	**100**	**100**
$JPEG_7$	0	0	0	0	**100**	**100**
$JPEG_8$	0	0	0	0	**100**	**100**
$JPEG_9$	0	0	0	0	**100**	**100**
$JPEG_{10}$	0	0	0	0	**100**	**100**
$JPEG_{11}$	0	0	0	0	**100**	**100**
$JPEG_{12}$	0	0	0	0	**100**	**100**

A classification accuracy of 100% is encouraging, however in automotive imaging systems levels of compression corresponding to the worst levels (10-12) of distortion in our dataset are rare, and only likely in failure conditions or potentially when driving in low light conditions. Nevertheless, classification accuracy remains high for realistic levels of automotive compression (levels 4-6). Next, the influence of image degradations on the performance of the pedestrian detection algorithm were investigated. Figure 4.5 illustrates the degradation of performance of the RAW-classifier pedestrian detection algorithm on AWGN corrupted and JPEG compressed images in the data set. It is not surprising that the decrease in pedestrian detection performance with increasing levels of degradation is mostly monotonic. However, it is also interesting to note that the results from Chapter 3 have been validated on a much larger pedestrian data set consisting of more realistic automotive scenes.



(a)



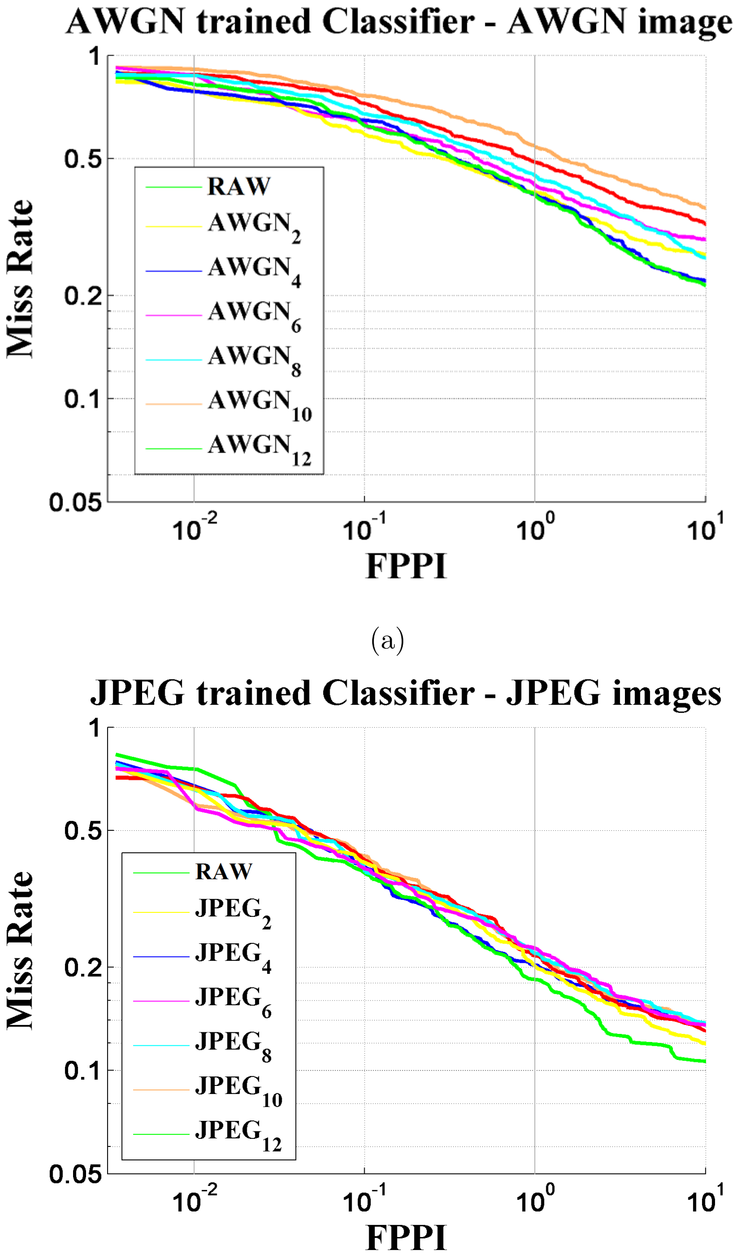
(b)

Figure 4.5: LAMRs of pedestrian detection performance on the AWGN degraded (a) and JPEG compressed (b) versions of the database. It is observed that for a particular range of quality parameter, a small increase in image quality can yield a large increase in detection performance.

4.4 Classifier Retraining

Classifier retraining has been shown to be a significant factor affecting overall detection performance [123]. The influence of classifier retraining on the performance of the ICF pedestrian detection algorithm was investigated on this larger data set. The detection algorithm was trained multiple times with distorted versions of the INRIA and Caltech training data sets and tested on the non-overlapping test data sets. From Figure 4.6, it can be seen that performance improvements can be gained by training to specific distortion types and levels. For example, Figure 4.6(a) reports the performance of a classifier trained with low quality AWGN corrupted images. In comparison to Figure 4.5, it can be seen that this classifier improves detection performance significantly on images of similar quality and distortion type to the training set. For example, the LAMR of the RAW trained detection algorithm on AWGN level 10 is 83.95%, compared with a LAMR of 58.49% using the AWGN trained classifier. Similar improvements in quality are shown in Figure 4.6(b) for JPEG compressed images. Details of the training samples used for each classifier are listed in appendix A.

Although these classifiers optimize performance for specific distortion types and quality levels, they are not the optimal solution for the entire data set. Again, an optimized single classifier was obtained by training classifiers on images with a wide range of quality degradations. However the single classifier offered only a modest performance improvement over the RAW classifier on this larger data set. It was attempted to improve performance further by increasing the size of the training set. Classifiers were trained with all degraded versions of the INRIA and Caltech train-



(a)

(b)

Figure 4.6: Performance of a classifier trained with poor quality AWGN images (a) on all noise corrupted images in the data set, and performance of a classifier trained with highly compressed JPEG images (b) on all JPEG compressed images.

ing sets, consisting of over 18,000 and 100,000 images respectively in an attempt to increase the detection algorithm's robustness to image degradations. Again, little resulting improvement in overall detection performance was found.

The proposed multi-classifier framework is well suited to the wide variety of image quality levels that can occur in automotive visual communication channels and hence ought to significantly improve overall detection performance. For each distortion type and NIQE bin width described in Section 4.2.3, a detection classifier was trained with images from this specific quality level. The proposed categorization methodology was then used to determine the appropriate classifier for each image in the test data set. The results, shown in Figure 4.7, illustrate that the multi-classifier detection framework offers significant improvement over RAW, optimal, and robust single classifier performance. In Tables 4.3 and 4.4 the LAMRs for the top performing single classifiers on both INRIA and Caltech data sets are reported, as well as the best performing distortion specific classifiers, which, for both JPEG compressed and AWGN corrupted images, were achieved by training the detection classifier with images ranging from 6 to 10 on the NIQE quality scale. The proposed multi-classifier framework offers higher detection accuracy across both data sets and also for each specific distortion, providing higher tolerance against degraded images than any of the single classifiers. A Matlab script to evaluate detections against ground truth is included in appendix B.

The performance of the proposed algorithm was further evaluated on real video data that included impairments such as sensor noise and compression artifacts. All outdoor video sequences of pedestrians from the CD 2014 data set [127] were selected

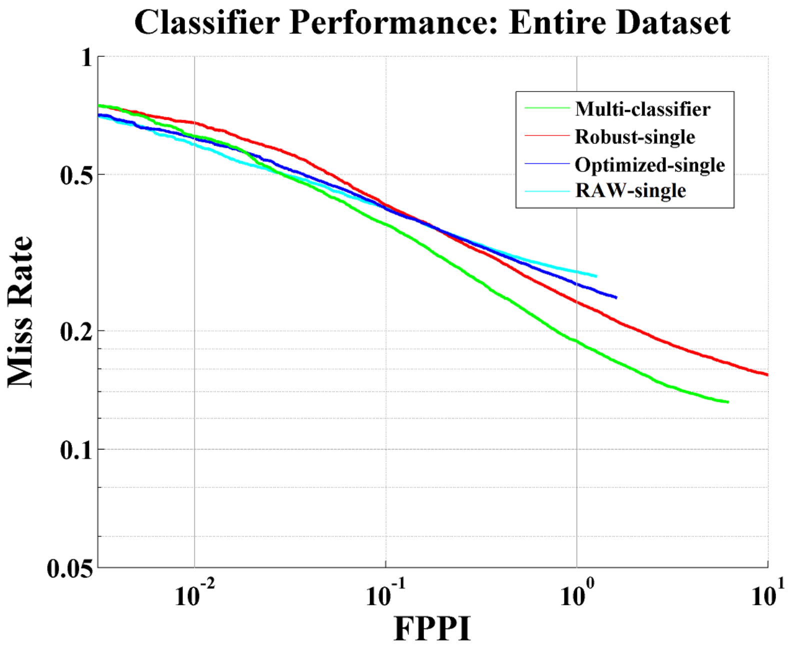


Figure 4.7: The performance of the proposed multi-classifier "quality aware" detection framework offers significant improvement over single classifier detection models on the INRIA data set.

as a real video benchmark. Since the videos in the CD data set were obtained with different cameras ranging in spatial resolution and compression parameters, the levels of image quality vary from approximately 3 to 8 on the NIQE scale. The results are summarised in Table 4.5 and Figure 4.8. The results show that the multi-classifier framework is also the best performing algorithm on this data set, offering performance gains of 3%-7% over the RAW trained classifier. In practical terms, this equates to between 3%-7% less missed detections at a given false positive rate, thus improving system performance and potentially reducing road fatalities.

Table 4.3: Detection Classifier Performance (LAMR) on the INRIA data set

Classifier	**JPEG**	**AWGN**	**Full data set**
RAW	28.25	52.59	41.47
Optimal	30.31	50.34	40.95
Robust	27.57	49.54	40.71
***JPEG**	26.54	-	-
***AWGN**	-	48.67	-
Proposed	**25.00**	**42.82**	**35.95**

Table 4.4: Detection Classifier Performance (LAMR) on the Caltech data set

Classifier	**JPEG**	**AWGN**	**Full data set**
RAW	66.87	86.38	77.84
Optimal	65.43	87.91	76.02
Robust	64.06	85.31	76.47
***JPEG**	63.70	-	-
***AWGN**	-	83.77	-
Proposed	**61.81**	**78.14**	**72.44**

Note: Detection performance is reported as log-average miss rates (LAMR) where lower scores represent higher performance. *JPEG and AWGN-specific classifiers are only evaluated on their respective subsets of the image data set.

Table 4.5: Detection Classifier Performance on the ChangeDetection data set

Classifier	**LAMR**
RAW	21.82
Robust	21.41
JPEG	27.58
AWGN	33.29
Proposed	**17.51**

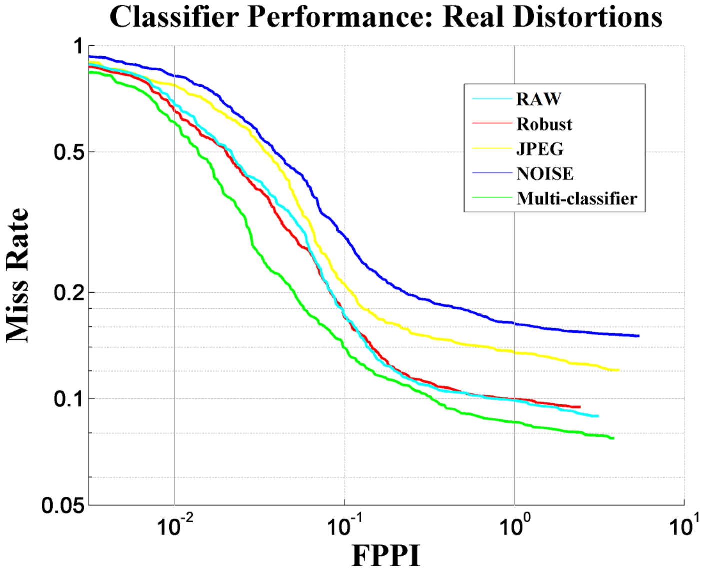


Figure 4.8: Relative performance of proposed multi-classifier detection framework, compared to noise-specific classifiers. The multi-classifier "quality-aware" detection framework is the best performing algorithm (in terms of LAMR) on real distortions from the ChangeDetection data set.

4.4.1 Computational Complexity

Since the features extracted for distortion classification are transform-free, the multi-classifier framework has relatively low computational complexity. In Table 4.6 the time taken (in seconds) to extract the NSS features from an image of resolution 640 × 480 pixels on a 3.3GHz PC with 8 GB of RAM operating on a single core is listed. Unoptimized MATLAB® code was used for all algorithms to ensure a fair comparison. As Table 4.6 demonstrates, the NIQE feature extraction stage is much more efficient than transform based methods such as BLIINDS or DIIVINE. This suggests that the multi-classifier method can be readily incorporated into real-time systems as typically, embedded automotive image processing algorithms are written in highly optimized C++ code, which offers substantial run-time improvement over MATLAB implementations.

Table 4.6: Complexity Analysis of NSS features

Algorithm	**Time(seconds)**
DIIVINE	82.42
BLIINDS-II	40.41
NIQE	**0.24**

4.5 Discussion and Conclusions

In this chapter a distorted image database with known distortion types and levels has been created. It has been shown that for particular levels of JPEG compressed images, a small improvement in quality can lead to a significant increase in the performance of a pedestrian detection algorithm. This is an important result, since in most automotive vision systems, the effect of such quality degradations can often be mitigated by controlling the compression ratio. The impact of classifier retraining on the performance of the ICF pedestrian detection algorithm has been verified on a large image data set. The results illustrate that detection performance is optimized on distorted images by training a classifier specifically for distortion level and type. These results have been used along with spatial NSS features to develop a method for multi-classifier pedestrian detection that can be used in real-time (online mode), and does not require reference images. The results of the experiments show that the multi-classifier approach offers higher detection accuracy on distorted images than single classifier performance. However in this study, JPEG and AWGN compression are considered separately. In automotive applications, a noisy image is passed through the imaging pipeline of a camera and then compressed. The algorithm was further evaluated on such a general case of "real-world" images, again offering higher detection accuracy. Hence the proposed framework represents an improvement on approaches based on single classifiers, as used in most of the existing literature. Furthermore, the image distortion categorization framework operates directly on multi-scale spatial pixel data and hence is computationally efficient and suitable for real-time implementation [131].

A possible disadvantage of the proposed framework is an increase in complexity as additional distortion types, such as motion and focus blur are added to the framework. As more distortion types are included it would also become more difficult to ensure high classification accuracy. This would be an interesting area for further research. These challenges notwithstanding, a key advantage of the proposed detection framework is its high adaptability, since alternative statistical models can be derived to suit the needs of particular image capture techniques.

The next chapter revisits the issue of perceptual quality assessment in the context of video transmission, and examines the influence of saliency on perceptual assessment of video corrupted by transmission errors.

Chapter 5

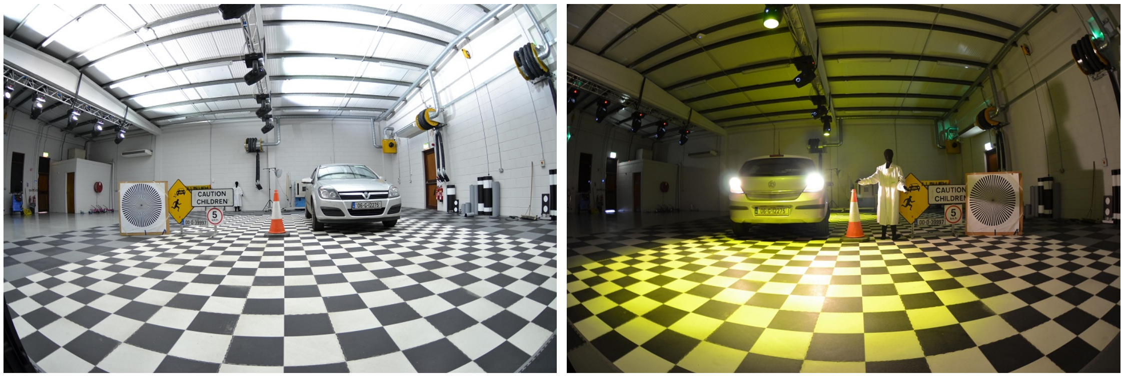
Evaluating the Influence of Saliency on Perceptual Quality in Automotive Vision Systems

5.1 Introduction

As noted previously, the Quality of Experience (QoE) concept for automotive video differs substantially from the traditional QoE concept for broadcasting or other consumer-related applications. In the automotive environment the subjective satisfaction of the user is related to achieving a particular task related to driving [132]. Borji *et. al* [133], describe visual saliency as factors that render certain image regions more conspicuous than others, for example image regions with different features from their surroundings. In automotive applications, the driver's attention is influenced heavily by salient objects in a scene such as a pedestrian or oncoming vehicle, which serves to change the manner in which drivers perceive image and video quality.

This chapter returns to the issue of perceptual quality, and in particular the relationship between perceptual quality and visual saliency, and whether a quantitative relationship can be established between the two.

In order to ascertain "ground truth" opinions on perceptual quality, image and video quality data sets with appropriate scenarios were generated. The details of each data set and subsequent subjective quality tests that were carried out are described in this chapter. In particular, this chapter investigates how distortions, such as blurring and packet losses, which are inherent to data acquisition in automotive vision systems, influence visual saliency, and hence perceptual quality. In Section 5.2, the characteristics of the image data set used for the first set of experiments that consider the effect of blurring are described. In Section 5.3, the subjective tests completed are described and some insights into the characteristics of the distortions used in the data set are considered, including loss of sharpness caused by radial lens distortion. A no-reference metric based on blur is proposed in Section 5.4 as a means of predicting the perceptual quality of fish-eye to rectilinear transformed automotive images. Since packet loss is a form of temporal distortion, it is best evaluated using video data. In Section 5.5 a new automotive video quality data set containing packet losses is described. A subjective video quality experiment carried out on this video set is described in Section 5.6. Eye tracking data were collected during the experiment to monitor each subject's viewing behaviour. Finally, in Section 5.7, a no-reference image quality metric for video corrupted with packet losses is proposed.



(a) (b)

Figure 5.1: Two diorama scenes from vision laboratory: scene (a) is illuminated evenly with ambient white light of 86 lux; Scene (b) simulates street light illumination.

5.2 Image Data Set

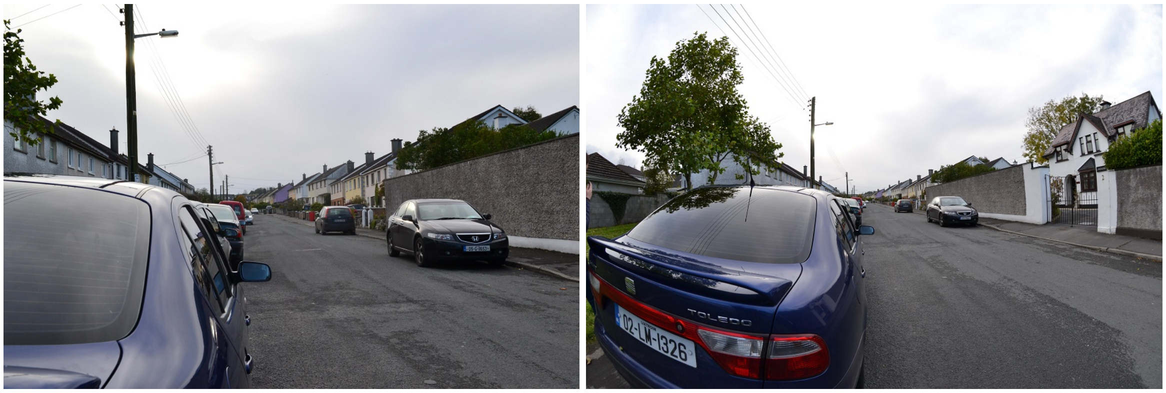
The first data set developed was an image data set that consisted of 16 automotive scenes. The scenes were chosen to span a range of typical automotive environments. In particular, 6 indoor diorama scenes were captured in a purpose-built industrial vision laboratory under controlled lighting conditions. The vision laboratory was designed such that ambient light could be measured and controlled in all areas of the scene. For example Figure 5.1(a) has a uniform illumination of 86 lux, which is representative of a fully overcast morning or evening, while the scene in Figure 5.1(b) is illuminated by two sodium lights providing an average luminance of 15 lux conforming to the recommendations for night-time lighting levels for roadways and sidewalks [134].

Accounting for variations in lighting conditions is important since light is known to affect human perception. For example, although the HVS is capable of adapting to a wide range of lighting conditions, high ambient light levels decrease its sensitivity to small variations in contrast [135].

Other lighting conditions in the data set include a typical night-time driving scene illuminated only by automobile headlamps, as well as a twilight driving scene with an average illumination of 30 lux. The diorama scenes in the data set are complimented by 10 real world outdoor automotive scenes captured in daylight driving conditions. Both indoor and outdoor scenes were captured with both a standard (24mm) Nikon lens and a 10.5mm Nikon fish-eye lens. An example of each image type is given in Figure 5.2. The 10.5mm fish-eye lens has a horizontal Field Of View (FOV) of 138 degrees which is consistent with that of the forward facing cameras commonly used in the automotive industry. The 24mm lens has a higher spatial resolution and hence would be expected to produce images of higher perceptual quality. However, the FOV of a standard rectilinear lens is insufficient for most automotive applications. Nevertheless, images taken with this camera provide a benchmark representing high perceptual quality.

As mentioned in Chapter 1, fish-eye lenses are problematic in automotive vision systems since the radial distortion necessarily introduced by such lenses alters the perspective of the image, making it difficult for drivers to accurately judge distance. For this reason, fish-eye lenses are typically converted to rectilinear images before being displayed to the driver.

In order to transform the 10.5mm fish-eye images, the approach described by Scaramuzza *et. al* [136] was utilized. The method described in that paper requires a calibration step, in which a few pictures of a planar pattern of known geometry

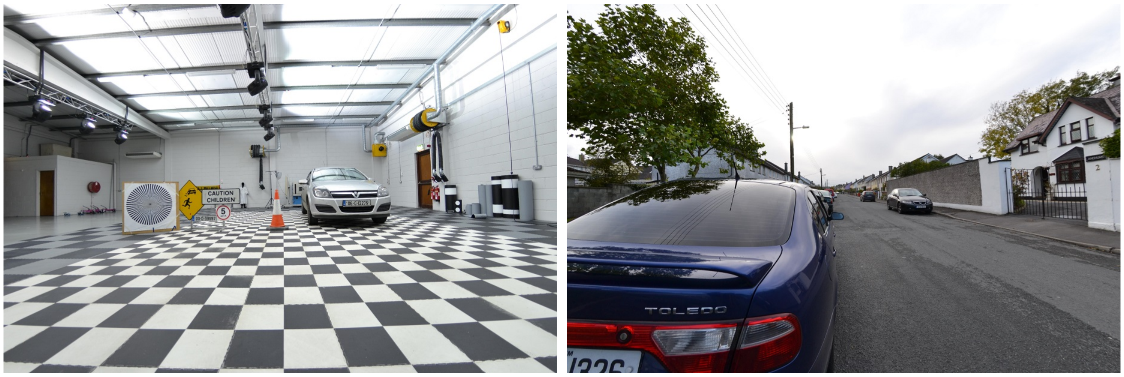


(a) (b)

Figure 5.2: A Nikon 24mm image (a); and corresponding 10.5mm fish-eye image (b).

are taken at different locations in the image. A Taylor series expansion is then used to model the image formation function, which projects a 3D real point onto a pixel on the image plane. The coefficients of the Taylor series expansion are estimated by solving a two-step least squares linear minimization problem. The order of the series is then determined by minimising the re-projection error of the calibrated points. This method was chosen for a number of reasons. Once the images are taken, the calibration procedure is very fast and entirely automatic. Moreover, the Taylor series approximation to the lens was found to be very accurate. In the experiments described in this chapter, a checkerboard image was used to calibrate the 10.5mm Nikon fish-eye lens. After calibration refinement, a Taylor series expansion of order 4 was computed to approximate the image formation function. On images of resolution 4608 × 3072 pixels, the average re-projection error is only 1.51 pixels with a standard deviation of ± 0.74 pixels.

Some examples of the rectilinearly transformed images are shown in Figure 5.3.



(a)

(b)

Figure 5.3: 10.5mm fish-eye images corrected to rectilinear form using a Taylor series expansion. (a) Image of Figure 5.1; (b) Image of Figure 5.2.

The scene in Figure 5.3(a) is the same as that used in Figure 5.1, while that in Figure 5.3(b) is the same as that in Figure 5.2. Notice that the FOV in Figure 5.3(b) is far wider than that of the corresponding 24mm image in Figure 5.2(a). Curved beams in the fish-eye images of Figure 5.1(a) are mapped to straight lines in Figure 5.3(a). The complete image data set therefore contains three types of image, namely 24mm rectilinear images, 10.5mm fish-eye images and 10.5mm corrected-to-rectilinear images.

5.3 Subjective Image Quality Studies

In order to assess the perceptual quality of the images in the data set, two Absolute Category Rating with Hidden Reference (ACR-HR) [16] studies were conducted at the National University of Ireland, Galway (NUI Galway) over the course of three weeks. The subject pool consisted of 35 (mostly post-graduate) students from NUI

Galway who had no previous exposure to subjective image tests. The subjects were a mix of males and females with a male majority. A verbal confirmation of visual acuity was obtained from each student prior to each study and all subjects were tested for colour blindness. Each study involved a single viewing session lasting under 20 minutes, in order to minimise viewer fatigue [14]. In the first study, subjects were asked to rate the usefulness of the images for the purpose of driver assistance. De Ridder and Endrikhovski describe "usefulness" as the degree of apparent suitability of an image with respect to a specific task [137]. For the second study, the same set of images was used, but on this occasion the participants were asked simply to rate the perceptual quality of each image. The average testing time per subject was approximately 16 minutes.

An informal after-study questionnaire indicated that viewers did not experience any fatigue during the course of the study.

Each study began with a short training session during which the subject was presented with six images chosen to span the range of distortions contained in the data set. The images used in the training session differed from those used in the actual study but were of a similar nature. The studies used a data set of 48 images shown in random order. Furthermore, the order was randomized for each subject and care was taken to ensure that two consecutive images did not correspond to the same reference image to minimise memory effects. Images were displayed on a 24 inch Dell ST2421L monitor with a screen resolution of 1920 × 1080 pixels at a viewing distance of 1 metre. The study took place in a dedicated viewing room with low background illumination as per recommendations in [14]. A subject rejection procedure, described in [14]

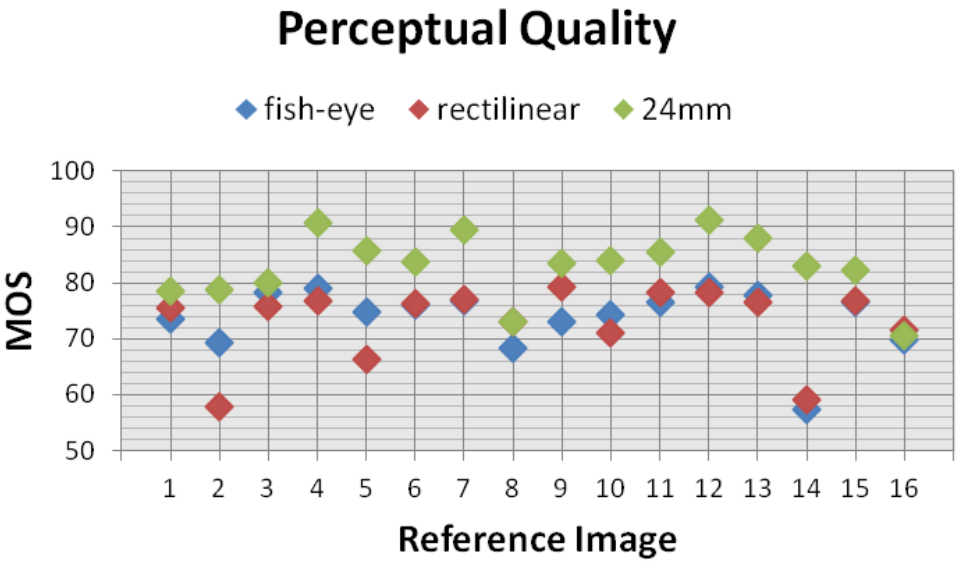


Figure 5.4: MOS scores evaluating the perceptual quality of images in the data set. Note that the reference data set consists of 16 images, each of which is represented three times in fish-eye, 24mm lens and rectilinear transformed versions.

was carried out that rejected one subject from each study. The remaining subjective scores were then averaged across subjects to obtain Mean Opinion Scores (MOS) for each image.

Figures 5.4 and 5.5 illustrate the MOS scores for the images in the database, for usefulness and perceptual quality. Images taken with the 24mm lens have the highest pixel resolution and also the highest perceptual quality, however the "usefulness" of these images is far lower than that of both the fish-eye and corrected images due to the much narrower field of view. Perceptual MOS scores for most rectilinear images were comparable with those of the fish-eye images. Average MOS scores after conversion to percentage values were 83.01, 73.78 and 73.09 for the 24mm, fish-eye and corrected

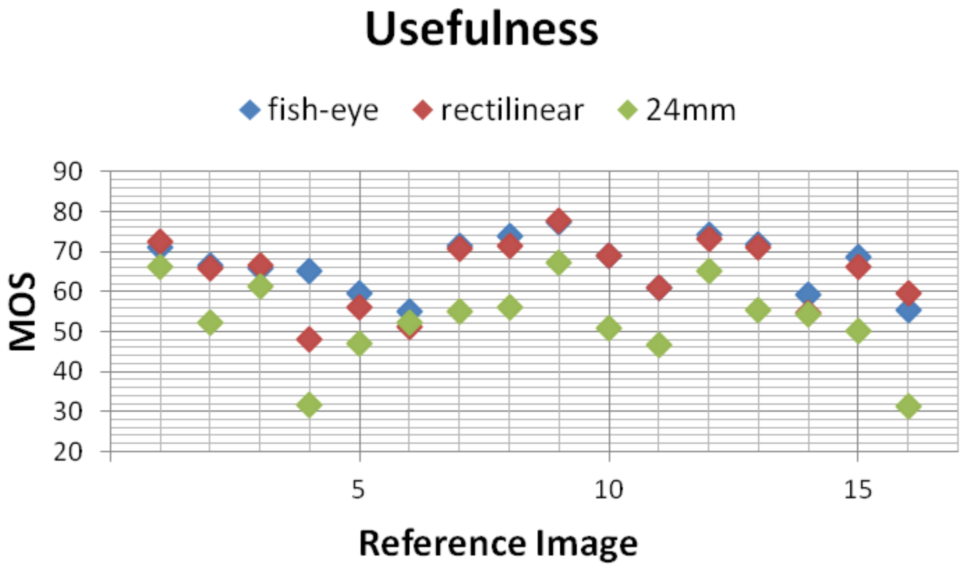
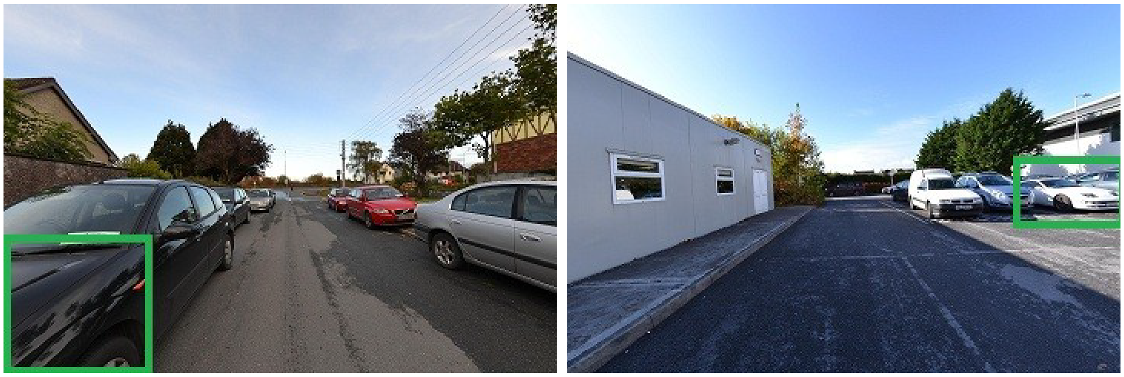


Figure 5.5: MOS scores rating the "usefulness" of images in the data set for driver assistance applications.

rectilinear images respectively. Additionally, results from the subjective study showed that the corrected rectilinear images were almost equally as "useful" for automotive driver assistance as the fish-eye images which have only a slightly wider field of view (average MOS scores in percentage terms were 64.7 and 66.5 respectively, while fields of view are 138 and 137 degrees respectively). Both of these images achieved far greater MOS scores than the 24mm images (whose average MOS was 52.7) when evaluated for "usefulness".

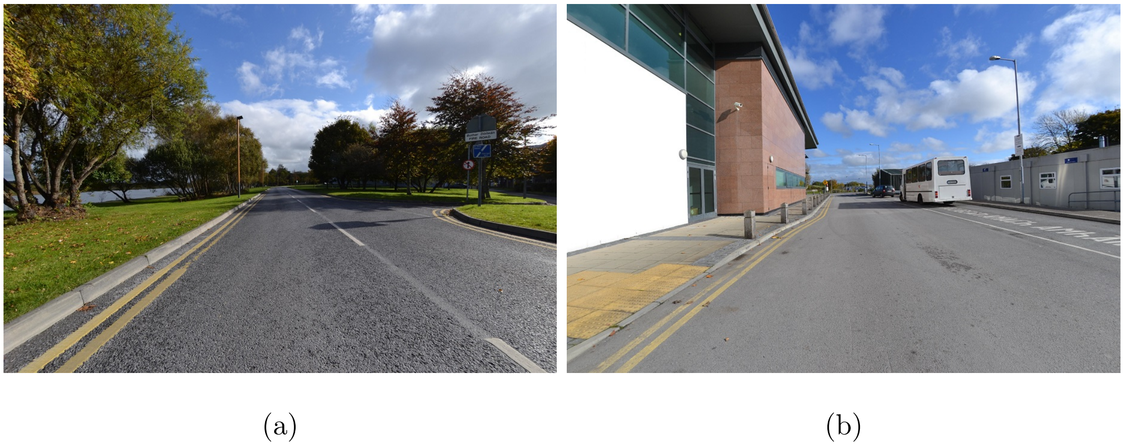
The experiments also showed that the location of salient objects has a bearing on the perceptual quality of a distorted image. For example, images "2" and "5" in the data set (Shown in Figure 5.6(a) and (b) respectively) exhibit the highest drop in perceptual quality relative to the fish-eye image after rectilinear correction.



(a) (b)

Figure 5.6: Salient objects at the edge of transformed images suffer from distortions. These distortions contribute to poor perceptual quality scores.

Notice that in both of these images an object of interest (a vehicle) is present at the edge of the image where the error in the polynomial approximation to the lens is greatest. These images suggest that naturalness plays an important role in perceptual image quality, but also highlight the difficulty of evaluating perceptual quality in this environment since the location of salient features in an image can affect a viewer's opinion of quality. There is no such drop in perceptual quality when salient objects are absent from the peripheral areas of images, such as shown for example in the images in Figure 5.7.



(a) (b)

Figure 5.7: Transformed images that contain no salient objects in peripheral regions do not suffer from a loss in perceptual quality.

Since fish-eye to rectilinear transformation results in a loss of spatial resolution as a function of distance from the image centre, a noticeable decrease in sharpness is evident in the peripheral areas of the transformed rectilinear images. This drop in sharpness is illustrated in Figure 5.8. A block of pixels in the centre of the image is compared to a similar block at the periphery of the image. In this particular example two 64×64 pixel blocks were chosen from the wall in the background. Both sets of pixels are of a similar object and hence should have similar properties with regard to colour and texture. However, a significant difference in high frequency components is evident from the Fourier spectra shown as inserts in the Figure. The Fourier spectrum from the centre of the image contains much more high frequency energy than that obtained from the edge of the scene. The drop in high frequency components represents a loss of fine detail in the image. The loss of sharpness is further illustrated by analysing a horizontal strip of each image in the data set (in the automotive environment a horizontal strip contains much more pertinent scene information than a vertical one, since a horizontal strip spans the width of the road ahead, where objects of interest are most likely to be located). The decrease in high frequency energy (calculated from spectral analysis) as a function of pixel distance from the optical centre of the lens is shown in Figure 5.9. The black line in the Figure represents the average high frequency energy across all fish-eye images at a particular distance from the image centre. Although high frequency energy depends on the local image content it is also evident that energy tends to decrease with distance from the image centre.

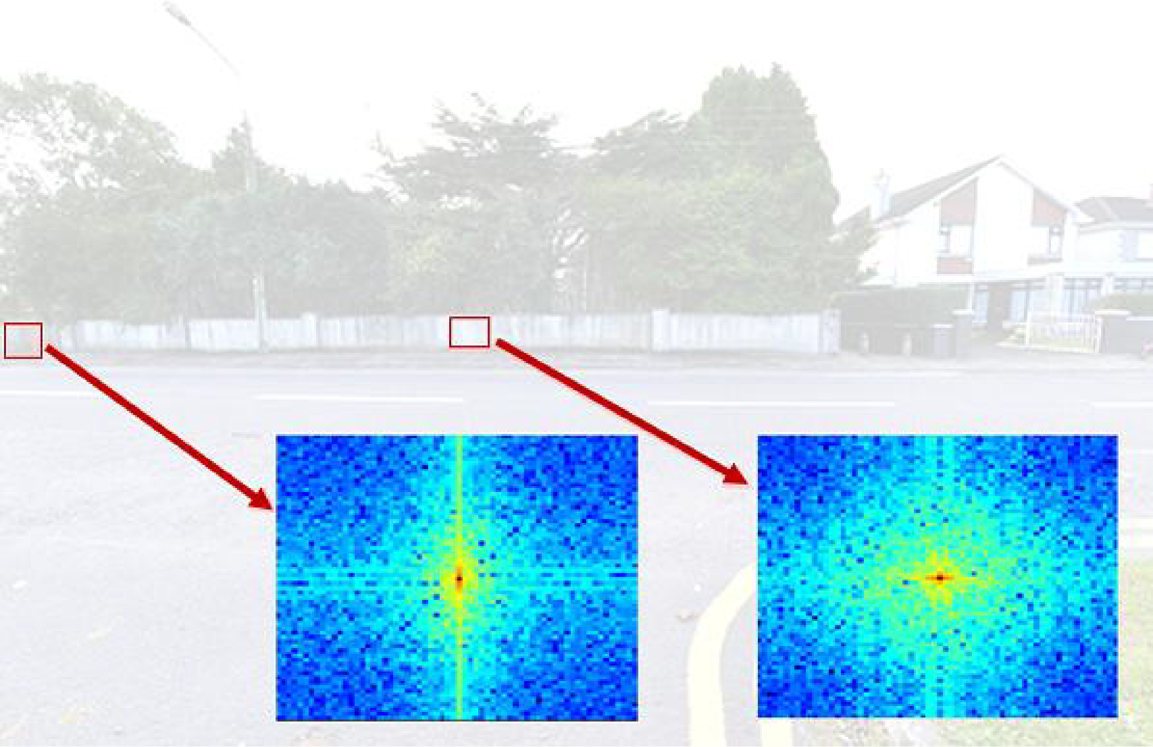


Figure 5.8: A loss of resolution is characterised by less high frequency energy in the Fourier spectrum.

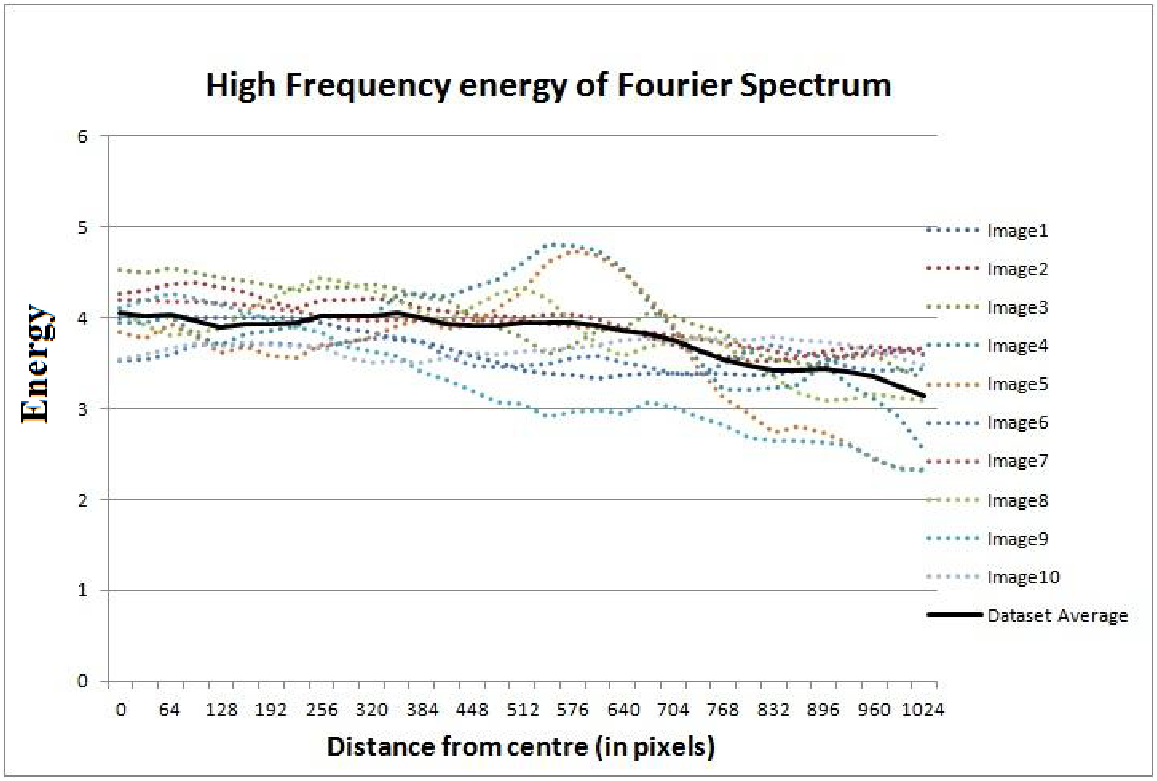
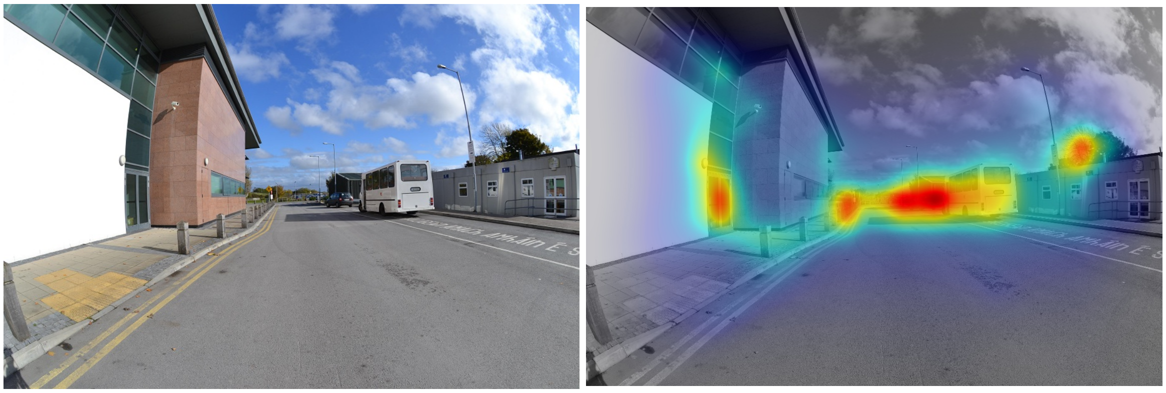


Figure 5.9: High Frequency energy loss in the Fourier Spectrum is used to estimate local image sharpness. Images become more blurred as distance from the centre of the image increases.

5.4 No-Reference Blur Metric

As noted in the introduction, blurring is an important form of distortion affecting image quality that is a common component of typical automotive image acquisition systems. Since image blur is non-uniform throughout automotive images, different regions of an image may have significantly different perceptual quality. For this reason it is necessary to consider which regions of the image may be more salient, or contain more pertinent visual information to the driver. Such image regions are more likely to attract the attention of the driver and hence more strongly influence his or her perception of image quality. To account for differences in image quality caused by different levels of saliency, the local image quality score of an image region can be weighted by its predicted visual importance. Such saliency weighted, no-reference distortion metrics have been proposed, for example in [68, 138, 139]. These metrics typically outperform their non-weighted equivalences since regions of visual importance are weighted more highly than regions with little structural content. The leading models of visual saliency [140, 141], including the well-known benchmark saliency model developed by Itti, Koch and Niebur [142] may be organized into three stages [141]. In Stage 1, feature vectors are extracted at locations over the image plane. During Stage 2 an "activation map" (or maps) is formed using the feature vectors. Finally, in Stage 3 the activation map (or maps, followed by a combination of the maps into a single map) is normalized. These models are based on the properties of the early HVS and in particular the theory of feature integration (also known as bottom-up visual processing). This theory suggests that when the HVS perceives a stimulus, features are registered early and in parallel, while objects are identified separately at a later stage of processing.

In [141] a state-of-the-art bottom-up saliency model based on graph contributions termed Graph Based Visual Saliency (GBVS) was proposed. In a comparison of GBVS [141] against existing algorithms on a data set of images of natural environments the model compared favourably to the more traditional Itti-Koch-Niebur algorithm achieving 98% correlation with human saliency estimates, compared to a correlation of only 84% between the Itti-Koch-Niebur model and human saliency. An example of the GBVS algorithm working on the data set used here is illustrated in



(a) (b)

Figure 5.10: The “hot” regions on the image (b) correspond to areas of predicted high visual saliency.

Figure 5.10. The “heat map” corresponds to image areas of high visual saliency. In order to predict the perceptual quality of rectilinear transformed automotive images both the graph based and Itti-Koch-Niebur visual saliency maps were computed for each image in the data set. The Fourier energy of local image regions was then computed and weighted by the normalized saliency map. An image quality metric Q was then computed according to the following formula:

$$Q = \frac{1}{n} \sum_{ij} S_{ij} F_{ij} \tag{5.1}$$

where S_{ij} is the normalized local saliency weight (calculated according to either the GBVS or Itti-Koch-Niebur methods), F_{ij} is the local high frequency energy of the Fourier spectrum, and n is the number of local regions computed in each image.

The predicted quality values from the proposed algorithm were correlated against subjective perceptual MOS values from the subjective image study. The results are illustrated in Table 5.1. Since the GBVS map outperformed the Itti-Koch-Niebur map, this method was chosen to weight the local sharpness scores for the remainder of the analysis.

The saliency weighted algorithm is shown in Figure 5.11 to correlate closely with perceptual quality for daylight driving conditions. Saliency information increases the algorithm's correlation with MOS values from 0.7732 to 0.8268.

Four of the images in the data set were taken in night-time driving conditions. Correlation of metric score to perceptual quality for these images is poor, since poor luminance and contrast can mask the presence of image blur. Moreover, in low light conditions, the predominant image artifact affecting perceptual image quality is thermal noise on the image sensor as can be seen in Figure 5.12. Nevertheless, the results illustrate the importance of saliency in image quality evaluation. In the next section the effects of quality impairments on automotive video are considered, taking viewer saliency into account.

Table 5.1: Saliency Weighted Metric Performance

	Pearson Correlation Coefficient
GBVS Quality Metric	0.8268
Itti-Koch-Neibur Quality Metric	0.7942
Non-Weighted Quality Metric	0.7732

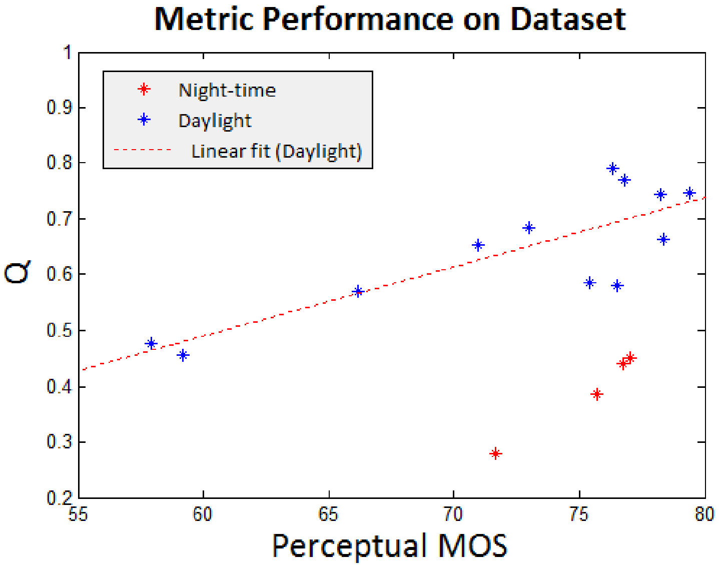


Figure 5.11: The performance of the proposed algorithm on the data set of transformed images. The x-axis represents perceptual quality as a percentage, while the y-axis represents the GBVS weighted quality metric.

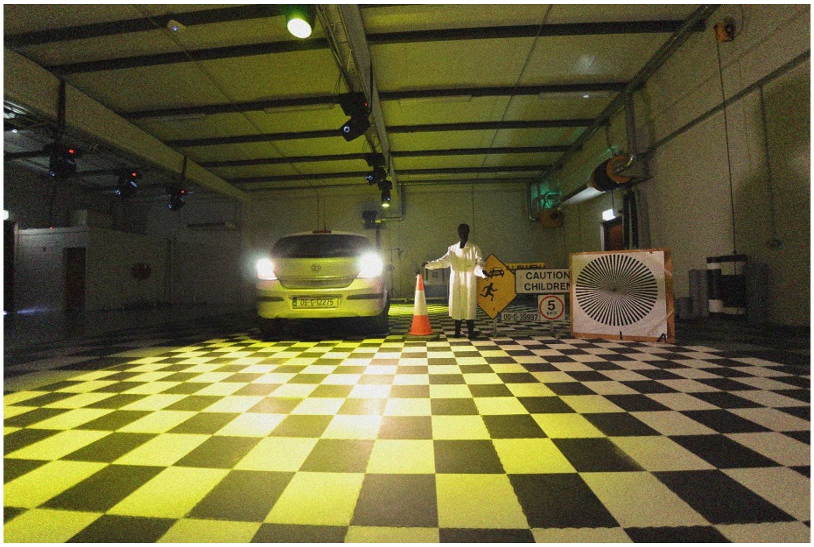


Figure 5.12: Example image taken in low light conditions. Thermal noise from the image sensor is the predominant distortion.

5.5 Video Quality Data Set

Although image quality data sets can be useful for establishing ground truth quality ratings for *spatial* lens degradations such as image blur, in order to evaluate *temporal* impairments it is necessary to consider video data. In typical automotive in-vehicle networks, video quality can be degraded due to packet losses, however the influence of such quality impairments on the QoE of the driver is not well understood. To investigate this impairment on quality, an automotive specific video quality database was created by collecting real video data from an in-vehicle, driver assistance system with a rear-facing camera. Ten reference video sequences were selected from the collected data and subjected to varying degrees of simulated packet losses in order to model a lossy automotive network. The impact of network impairments on the viewer's perception of video quality was then assessed by conducting a subjective test. Saliency data were collected from each viewer to further assess the influence of packet losses on visual attention.

5.5.1 Automotive Network Simulation Methodology

An automotive Ethernet network topology simulation was used to simulate realistic packet losses in the reference video sequences. Ethernet is fast becoming one of the most utilized and well researched technologies in automotive networking [143]. There are a number of factors that make Ethernet an appealing choice for in-vehicle networking. It is widely used outside of the automotive domain, unlike some current generation technologies, therefore it is a cost effective option for high volume manufacture. Additionally, its widespread deployment means that it is actively being developed and iterated upon to provide more functionality and higher bandwidth.

The network used in this experiment is based on the ns-3 simulation platform [144], which allows the introduction of real traffic streams to a simulated network, instead of relying on generic simulated traffic generators. Extensive work has been carried out by Tuohy et al. [145] in modelling in-vehicle automotive networks. In [145], a network simulation was developed to provide a platform for the preparation of packet loss-impaired video. Packet loss was introduced to video streams through the use of a mathematical error model on simulated Ethernet links connecting cameras to a video receiver node. This works by marking packets which pass through a link for dropping according to a mathematical model. Using a random number generator and a pre-configured percentage value, packets are dropped in bursts such that the overall occurrence of dropped packets is based on a random distribution, but also has a predictable percentage.

A top-down diagram of the network topology proposed by Tuohy *et al.* [145, 146] and used in these simulations is shown in Figure 5.13. The network models a triple star or daisy chain in-vehicle network, a topology that is common in next generation automotive networks [147, 148, 149]. Cameras were assumed to be the sources of the video of interest, while additional sources of data were added to model other sources of traffic that adds congestion to the network. The network contained the following elements:

Optical Cameras: The camera nodes were attached to Linux containers, which used a custom built video streaming application to send source videos of resolution 944 x 531 pixels at 25 frames per second across the simulated network. These samples were captured from cameras in a real world vehicle on public roads. The samples were transmitted across the network as uncompressed RGB image frames , in order to avoid the introduction of compression artifacts.

Infotainment and Miscellaneous Traffic: For infotainment and other traffic that is not part of the video data of interest (which could come from a 3G internet connection or wireless node within the vehicle), the flows were modelled as TCP streams.

CAN Gateway: The data rate of a CAN bus is limited by its length. Since a vehicle is a small space and CAN connections within are generally less than 10 meters in length, a bit rate of 400 kbps was assumed. A frame size of 20 bytes was used, which represents the maximum frame size allowable by the CAN bus standard.

Radar Device: The Radar sensor device contained in the simulation outputs data across a FlexRay gateway at 10 Mbps.

Samples were transmitted through the network at 4 different levels of impairment to model video corruption due to the noisy automotive environment.

On the receiver side, a standard repetition type insertion repair technique [150] was carried out on the received frames of video. While there exist a large number of different techniques for the mitigation of packet loss [151], such as interpolation, interleaving and retransmission, in an automotive scenario, the minimisation of delay between receipt of a frame and its display to the driver is extremely important. The repetition type insertion repair technique was chosen because it offers a combination of low computational cost, high speed and good subjective performance.

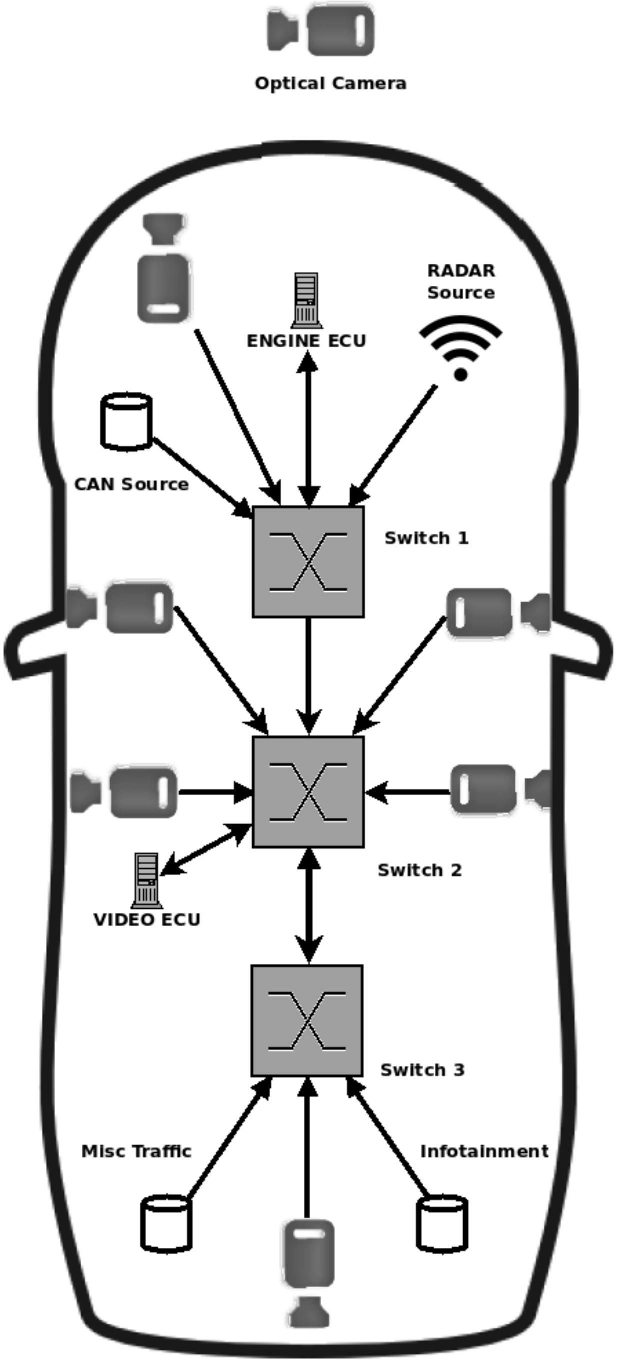


Figure 5.13: The network topology used to generate packet loss corrupted video

5.6 Subjective Video Quality Experiments

5.6.1 Data Set Design

The goal of this subjective experiment was to evaluate the effect of packet loss on the quality of user experience for automotive applications. The particular application under test was a rear-facing camera automotive display system, normally utilized for parking assistance systems. The reference video sequences used in the experiment were all of urban driving scenes captured from a fish-eye, in-vehicle, rear-facing camera at a frame rate of 25fps and cropped to an aspect ratio of 16:9.

Four levels of impairment (P1 - P4) were chosen to span a wide range of video quality. The percentages of packet loss for each level of impairment (P1 - P4) were 1%, 2.5%, 5% and 10% respectively. Similar levels of packet loss were used in [151, 152]. Although 5% and 10% levels of loss are high for ethernet networks, future automotive networks may employ inter-vehicular video transmission on networks more prone to such high losses. Ten reference video sequences were used for the subjective test varying from 8 to 24 seconds in duration. The average duration of the reference videos was approximately 16 seconds.

An example of a video frame extracted from the simulation after being subjected to 10% bursty packet loss can be seen in Figure 5.14. Large levels of impairment result in 'blocky' sections where movement has taken place between frames. These impairments are visually similar to the effects that are seen on other types of streaming video undergoing packet loss [151] [153].

Choosing appropriate sequences for a subjective quality test requires consideration

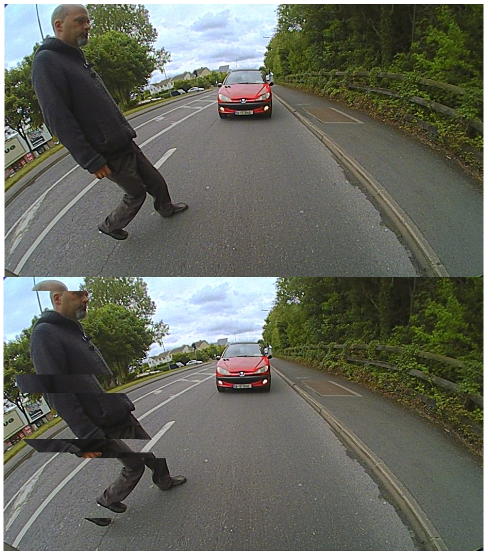


Figure 5.14: An input and extracted output frame from simulation after undergoing 10% packet loss.

of both the spatial and temporal content of each scene [154]. Automotive video quality tests should represent a wide range of spatial and temporal content in order to represent as much diversity as possible. Of course, it is impossible to include every conceivable automotive scene in a single subjective test, but increasing the diversity of the test set improves a test's accuracy. Including a wide variety of video content in the test set serves two purposes. First, it allows the packet-loss recovery algorithm to be tested rigorously. For example, scenes of low temporal activity may be relatively unaffected by the receiver-side insertion-repair technique due to little difference between two consecutive frames. On the other hand, a sequence with a pedestrian running across the road could be problematic for an insertion-repair error concealment algorithm since filling the missing video with prior content may cause part of the pedestrian to disappear. Second, it is well known from psychological vision science that the sensitivity of the HVS to video impairments varies with the spatial and temporal activity of the sequence [29]. Including a wide variety of spatial and temporal content in the test set enabled an evaluation of these effects on the visual attention of the viewer.

To simplify the task of quantifying scene complexity, objective measurements were used as recommended in [154]. The spatial perceptual information over the entire video (SI) is defined as:

$$SI = max_{frame}(std_{space}[Sobel(F_n)]) \tag{5.2}$$

where F_n is the luminance-only video frame at frame number n, *Sobel(X)* is the sobel filter operation on image X, max_{frame} is the maximum value in the video sequence, and the std_{space} is the standard deviation of all pixels in a frame.

The temporal perceptual information (TI) is defined as:

$$TI = max_{frame}(std_{space}[F_n - F_{n-1}]) \tag{5.3}$$

Both SI and TI are combined according to Fenimore et al. [155] to determine the complexity of the scene:

$$SI(F_n) = rms_{space}[Sobel(F_n)] \tag{5.4}$$

$$TI(F_n) = rms_{space}[F_n - F_{n-1}] \tag{5.5}$$

where, rms_{space} is the root mean square over all pixels in a frame. The complexity is then given as:

Table 5.2: Characteristics of Reference Sequences

Video Sequence	SI	TI	Complexity
1	86.98	14.43	3.10
2	91.46	13.10	3.08
3	118.91	15.65	3.28
4	117.11	6.85	2.90
5	118.49	3.16	2.57
6	91.40	18.13	3.23
7	105.59	6.28	2.82
8	88.04	6.73	2.78
9	110.17	14.99	3.22
10	107.90	7.84	2.92

$$C = log_{10} \sum_{1}^{n} \frac{SI(F_n) \times TI(F_n)}{n} \tag{5.6}$$

The spatial and temporal perceptual information of the reference sequences used in the study are reported in Table 5.2.

5.6.2 Subjective Test Methodology

The subjective test was performed using the Absolute Category Rating-Hidden Reference (ACR-HR) method [18]. This method was chosen since typical rear-view automotive vision systems utilize fish-eye lenses with fields-of-view of up to 190 degrees [23]. It has previously been mentioned that these lenses introduce radial distortion to the image and so it is possible that the subjective opinion of a viewer could be biased by such distortions. In the ACR-HR method, the reference sequences are included in the subjective test, but without being identified to the subject. The addition of the hidden references avoids the potential problem of reference sequences being given poor subjective scores due to radial distortion. Viewers therefore rate the reference sequences as they would any other test sequence. The quality scores are then reported as differential mean opinion scores (DMOS) given by equation 5.7.

$$DMOS = MOS(original) - MOS(degraded) \tag{5.7}$$

The selected test methodology is derived from the ITU-T recommendation P.910 [18]. Only non-expert viewers, as defined by [153] participated in the subjective tests. A total of 26 viewers undertook the study which involved a single session lasting approximately 25 minutes in order to minimise viewer fatigue. An informal after-study questionnaire indicated that viewers experienced little fatigue during the course of the study. Each study began with a short training session, during which the subjects were presented with video sequences chosen to span the range of impairments contained in the test data. The actual study consisted of 50 video sequences shown in random order. Furthermore, the order was randomized for each subject and care was taken to ensure that two consecutive sequences did not correspond to the same reference sequence in order to minimise memory effects. Subjects were instructed to watch the entire sequence before voting, receiving on-screen instructions as to when to vote. Subjective ratings were reported on the five-point scale: “Excellent”, “Good”, “Fair”, “Poor”, and “Bad” [18]. The study took place in a dedicated viewing room with low background illumination. Sequences were displayed on a 24 inch DELL ST2421L monitor with a screen resolution of 1920×1080 pixels, with test sequences centered, at their original resolution of 944×531 pixels. The background screen illumination was mid-grey, conforming to recommendations in [18]. The viewing distance was 70cm, or approximately 4 times the video height. A subject rejection procedure outlined in

[14] was carried out which rejected one subject. The remaining scores were averaged across subjects to obtain the DMOS for each video sequence.

5.6.3 Eye-Tracking Data

Eye-tracking data were recorded for each subject so that the most salient regions of each reference video sequence could be ascertained. Saliency data were also tracked on the degraded images so that differences in viewing behaviour on images corrupted with packet losses could be studied. In order to track and record each subject's eye movements, an Eyetribe tracker [156] was used. Unlike many infrared eye-tracking systems, the Eyetribe tracker does not require the viewer to use a rigid head rest, rather the viewer must only be located within the tracker's trackbox, which is defined as the volume of space wherein the subject can theoretically be reliably tracked by the system. Thus the subject's head movements were unrestricted for the duration of each subjective experiment, enabling a more realistic viewing environment. A twelve point calibration step was performed on each subject before beginning the test. The calibration step took less than one minute and ensured that the accuracy of the eye-tracking device was optimized for each subject. The pixel coordinates of each subject's gaze fixation were recorded for every frame of video viewed in the experiment.

5.6.4 Deriving a Saliency Map

A saliency map was derived from the spatial pattern of fixations in the eye tracking data according to [157]. Each subject's fixation location was recorded for each frame of video data. The data from each subject were then averaged to obtain an average

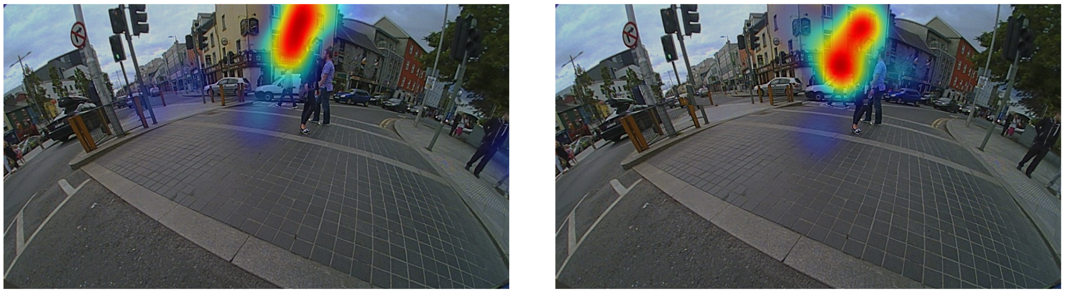


Figure 5.15: The saliency maps from a reference frame (a) and corresponding degraded frame (b).

fixation map ($FM_{(x,y)}$) for each sequence. A Gaussian distribution, the width of which approximates the size of the fovea (approximately 2° of visual angle) was then applied to each fixation point (x, y) in FM to obtain a mean saliency map (SM):

$$SM(k,l) = \sum_{i=1}^{T} exp[-\frac{(x_i - k)^2 + (y_i - l)^2}{\sigma^2}] \tag{5.8}$$

where $SM(k,l)$ indicates the saliency map for each given pixel (k,l) where $k \in [1, M]$ and $l \in [1, N]$. T is the total number of fixations, $(x_i,\, y_i)$ are the spatial coordinates of the i^{th} fixation and σ is the standard deviation of the Gaussian distribution. The intensity of the resulting saliency map was linearly normalized to the range [0,1]. Figure 5.15 illustrates an example SM derived from eye-tracking data obtained from a hidden reference image (Figure 5.15(a)) and, a SM of the same video frame corrupted with 10% packet loss (Figure 5.15(b)).

5.6.5 Analysis of the Subjective Test Results

In Figure 5.16, the DMOS scores are reported for each video sequence. Recall that the DMOS score is the difference between the reference and degraded MOS

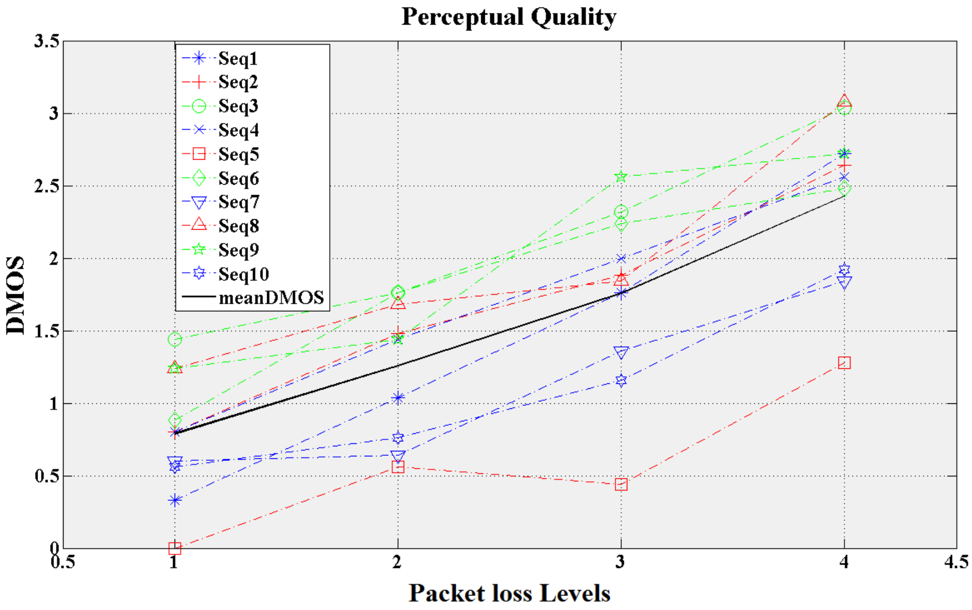


Figure 5.16: The DMOS scores from the subjective video test.

scores, hence a lower DMOS score represents a higher quality rating, thus reflecting lower perceptual impact of packet loss. It is observed that the decrease in perceptual quality with respect to increasing packet loss is largely monotonic, however there are a number of outliers which have significantly lower DMOS scores (higher perceptual quality ratings) than expected for the level of packet loss. In particular, the largest outlier (Sequence 5) is the sequence with the least movement, characterised by the sequence's TI score. The next two lowest DMOS scores also have low TI scores (Sequences 7 and 10). These sequences contain many instances of packet loss that were not noticed by the viewers due to similarities in consecutive frames. These results highlight the need to consider the temporal information in a video sequence in order to adequately assess quality impairments due to packet losses.

One of the key issues in studying perceptual quality is identifying factors which significantly alter visual attention. The Area Under the Curve (AUC) is a commonly used indicator to compare saliency maps [158]. It evaluates the area under the Receiver Operating Characteristic (ROC) which is found by plotting the false positive rate as a function of the true positive rate. Given a reference saliency map SM_{ref} and degraded saliency map SM_{deg}, the ROC is derived from the number of pixels labelled as salient in both SM_{ref} and SM_{deg} (true positives) versus the number of pixels labelled as salient in SM_{ref} that are not salient in SM_{deg} (false positives). A value of AUC = 1 indicates a perfect match, while a value of AUC = 0.5 indicates only a random match.

Eye tracking data and visual saliency are subject to high inter-observer variability [159]. Therefore in order to fairly evaluate the influence of packet loss on human attention it is necessary to calculate an AUC upper-bound, taking inter-observer variability into account. To determine an upper bound on similarity between saliency maps, the procedure adopted in [159] was followed to determine the Upper Empirical Similarity Limit (UESL). The UESL is defined as the maximum achievable similarity between the saliency maps derived from two groups of human observers under the same experimental conditions. For each reference video frame observed in the ACR-HR subjective experiment, subjects were divided into two randomly chosen groups, A and B, and their corresponding saliency maps SM_A and SM_B were calculated. The UESL was then computed as:

$$UESL = AUC(SM_A, SM_B) \tag{5.9}$$

The influence of packet loss on visual attention could then be defined by the normalized similarity (NS) which is the similarity between saliency maps obtained from both reference and degraded sequences, divided by the UESL:

$$NS = \frac{AUC(SM_{ref}, SM_{deg})}{UESL} \tag{5.10}$$

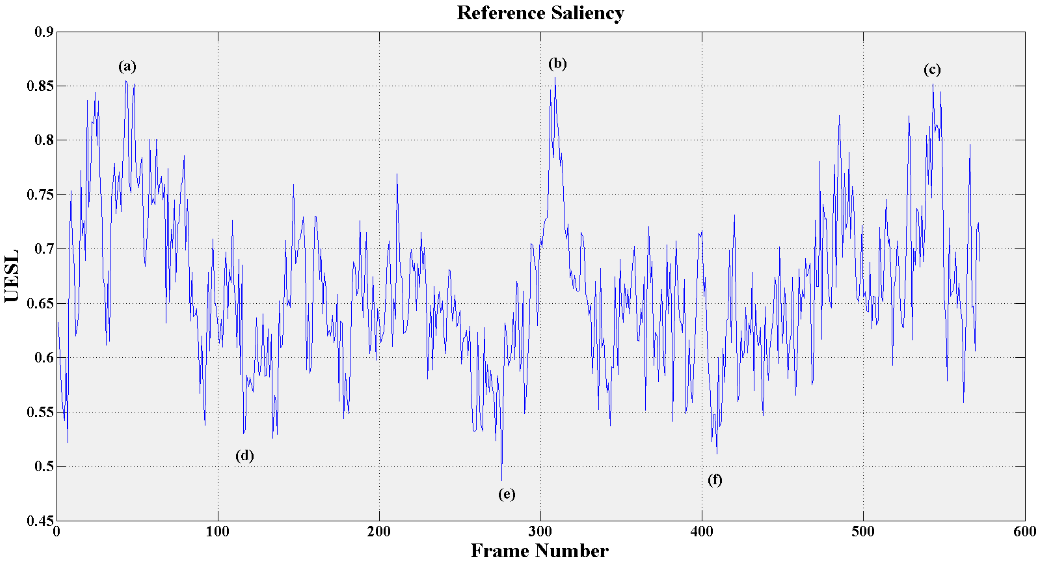
The normalized similarity thus gives a measure of the similarity between saliency maps obtained from video sequences with different quality levels, while taking inter-observer variability into account. Lower values of NS indicate lower similarity between saliency maps. It should be noted that because the limits are defined empirically, a value of NS greater than 1 is possible. The normalized similarity was calculated for every frame of every video sequence. A single NS score for each video sequence was obtained by averaging the NS scores over all frames in the sequence. These average NS values are shown in Table 5.3. The results show that increasing the level of packet loss had almost no effect on visual attention, despite significant differences in the MOS scores of different levels of packet loss. This is an interesting result and points to the attention of the viewer being more strongly influenced by task related (also called top-down) factors such as identifying potential dangers on the road, than so called bottom-up sensory cues which are more related to low-level vision. A study carried out in [158] found similar results in a free viewing task. To investigate this result further, the frame by frame saliency data of each video sequence were examined. The results show that there is little difference in viewer attention between the reference and corrupted video sequences. Figures 5.17 and 5.18 illustrate this point. Figure 5.17 highlights the similarity between saliency maps from a reference and corresponding corrupted video sequence (with 10% packet loss). The normalized similarity between the corrupted and reference saliency maps of this sequence is 0.9044, despite the corrupted sequence

Table 5.3: normalized Similarity of Degraded Saliency Maps, as a function of packet loss condition

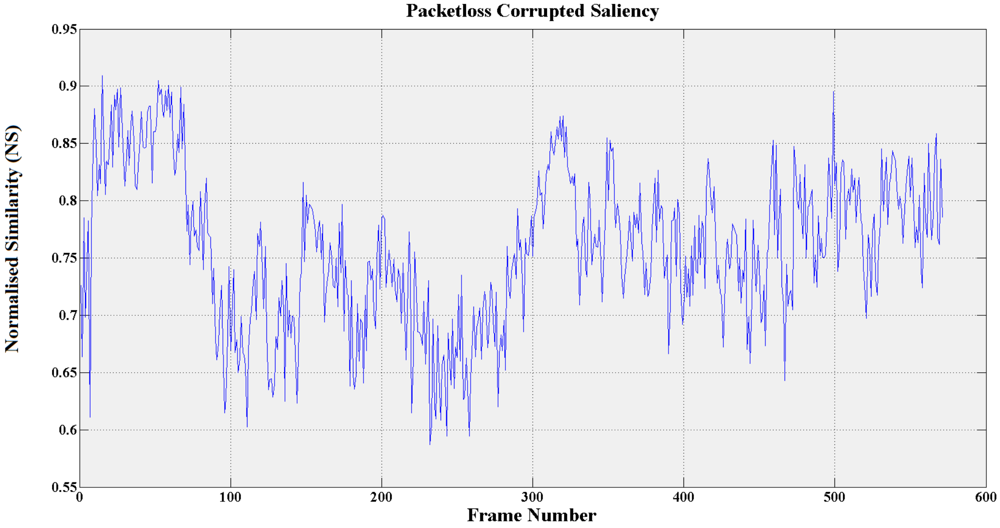
Sequence	P1	P2	P3	P4
1	0.8745	0.8686	0.8609	0.8706
2	0.9041	0.8876	0.9112	0.8944
3	0.8361	0.8543	0.8860	0.8667
4	0.8957	0.8806	0.9520	0.8766
5	0.9023	0.8949	0.9151	0.9192
6	0.8621	0.8678	0.8766	0.8468
7	0.8870	0.8974	0.9020	0.8593
8	0.8935	0.9029	0.9061	0.9183
9	0.8744	0.8763	0.8867	0.9044
10	0.8776	0.8859	0.8618	0.8821

having a DMOS score of over 2.5, indicating a high level of perceived degradation.

Peaks in UESL values in Figure 5.17(a) represent video frames where there is low inter-observer variability. These frames correspond to the images shown in Figure 5.18(a) - (c), where there is a single pedestrian in the frame and hence a focus of saliency. On the other hand, troughs in the UESL values from Figure 5.17(a) correspond to frames (d) - (f) in Figure 5.18, which contain multiple pedestrians and vehicles, and hence multiple regions of saliency. In general, it was found that the presence of pedestrians in a video sequence had the strongest influence on visual attention.



(a) The UESL computed over an entire reference sequence. Peaks in the UESL represent a high level of consistency between subjects' eye-tracking data.



(b) The normalized similarity (0.9044) of the saliency maps indicate that packet loss has little influence on visual attention, rather the content of the sequence determines the attention of the viewer.

Figure 5.17: Similarity Between Saliency Maps

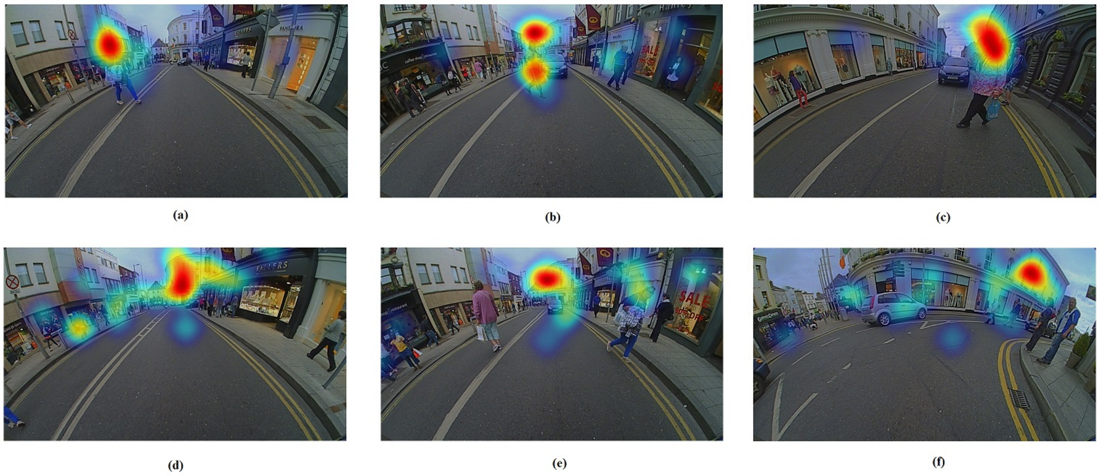


Figure 5.18: The frames with lowest inter-observer variability (a-c) correspond to frames with a single point of focus (a pedestrian), while frames with the highest inter-observer variability (d-f) contain multiple salient regions (many pedestrians or vehicles).

5.7 Development of a No-Reference Packet Loss Image Quality Metric

Packet loss degradation in the data set is neither spatially nor temporally uniform. Some areas of individual frames are degraded while the quality of other areas remains unchanged. Table 5.4 details the correlation between quality scores from 3 objective quality metrics and the subjective quality scores for the data set. The values are averaged across all test video sequences. In the case of the PSNR and SSIM metrics, the mean score across the entire sequence is used. Both PSNR and SSIM are full reference metrics, however, even with the use of a reference image they do not correlate well with human opinion scores for this type of packet loss. In fact, the recently-proposed no-reference video quality metric BLind Image Integrity Notator using DCT Statistics (video BLIINDS) proposed in [160] outperforms both full-reference algorithms. Nevertheless, the correlation with subjective MOS scores from the data set remains poor. Quality prediction could be improved by deriving a model of perceptual quality which incorporates the visibility of packet losses. This is considered in this section.

Due to the packet loss recovery model used (described in Sec. 5.5.1), the location of instances of packet loss could be determined by computing the correlation (inner product) between consecutive frames. Spatial regions where packet reconstruction has occurred are identical to the previous frame and hence have a correlation of 1, while non-corrupted image patches always have a correlation of less than 1. This is the case even for stationary consecutive frame sequences, since slight variations in pixel values always occur due to sensor noise at image capture. Although the method of locating instances of packet loss is specific to this recovery algorithm, alternative packet loss recovery methods typically introduce characteristic distortions in the video frame that can be distinguished from an uncorrupted frame; hence this general approach can be incorporated into alternative recovery algorithms with minor modifications.

Having found instances of packet losses based on correlation between consecutive frames, the goal was to categorise each instance as either salient or non-salient. For each frame containing a packet loss, the temporal difference between it and the previous frame was calculated according to:

$$TI_{frame} = (std_{space}[F_n - F_{n-1}]) \tag{5.11}$$

where the std_{space} is the standard deviation of all pixels in a frame.

It was hypothesised that packet losses are most salient if there is a large temporal difference between consecutive frames in the spatial region around a lost packet. Therefore, the temporal difference at the borders of each lost packet was measured. A border width of 10 pixels was chosen heuristically. Each local region around a lost packet was defined as salient only if the temporal information of that region exceeded a threshold of visibility chosen based on the quality scores observed from the subjective experiments.

The spatial texture of the lost regions was also considered, since regions with high texture have been shown to mask image degradations [28]. Entropy (H) is a statistical measure of randomness that has been widely used to characterise the texture of an image. The temporally salient packet loss regions were weighted by the spatial entropy in order to account for spatial texture, where the entropy of a probability density function $p(x)$ is defined as:

$$H = -\int p(x) \ln p(x) dx \tag{5.12}$$

The quality parameter Q for each frame is thus derived as the proportion of the frame degraded by visible packet losses:

$$Q = 1 - \frac{\sum_{i=1}^{n} S_{pl}}{n} \tag{5.13}$$

where n is the total number of pixels in the image matrix and S_{pl} are the computed visible packet losses in a video frame. The overall quality score for the sequence is computed as the average quality score across all frames of the sequence. By way of example, Figure 5.19 shows an image frame corrupted by two packet losses. The temporal difference between this frame and the previous frame in the sequence is

Table 5.4: Correlation of Quality metrics with MOS

Metric	PLCC
PSNR	0.3580
SSIM	0.3794
Video BLIINDS	0.4240

shown in Figure 5.19(b). The hypothesised “salient packet losses” are further weighted by the local entropy (c) to derive the proposed quality statistic (d) for each image frame. In this example, the most visible impairments due to packet losses occur near the back wheel of the van and on the pedestrians to the right of the video frame. These regions correspond to “highly salient packet losses” as predicted by the quality metric. The performance of the proposed quality metric was evaluated on ground truth MOS scores from the video database. Correlation between the derived quality model and human opinion across the entire test data set is 0.8211, which is significantly higher than the correlation for the other metrics shown in Table 5.4. The proposed quality model significantly outperforms both the PSNR and SSIM metric, as well as the more recent no-reference video BLIINDS algorithm, for video sequences degraded with packet loss.

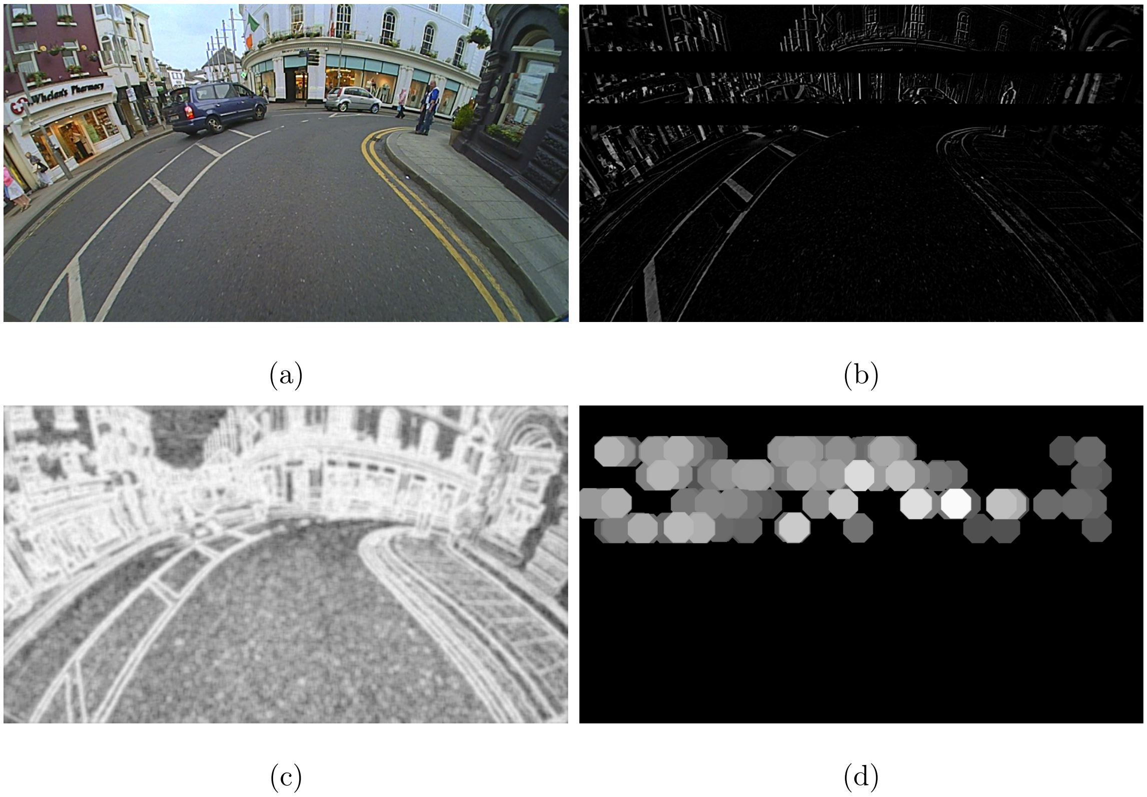


Figure 5.19: A frame with two lost packets (a), the temporal difference between frames (b), the entropy of the corrupted frame (c), and predicted salient packet loss (d).

5.8 Discussion and Conclusions

In this chapter a method for evaluating the perceptual quality of automotive images that have been converted from fish-eye to rectilinear images was first presented. It has also been demonstrated that the position of salient objects in a scene has a significant effect on perceptual quality. The use of regions of interest to weight objective metrics has been shown to increase correlation with subjective opinion scores. The influence of packet loss on QoE for automotive video has also been investigated.

A data set of automotive video sequences was created and transmitted through an automotive grade network simulation testbed with varying levels of packet loss. Subjective tests were conducted to obtain ground truth MOS and saliency data. The results indicate that packet losses do not significantly alter the visual attention of the viewer. This is an important result, since it suggests that the visual attention of a driver is more strongly influenced by top-down, task related factors such as watching the road ahead for potential hazards, than by sensory cues, which are more related to low-level vision. The MOS scores further suggest that packet losses in regions of high temporal activity are more salient than those in regions of low temporal activity. A no-reference model for evaluating the quality of video corrupted with packet losses was finally proposed. The results of the subjective experiment demonstrate that the proposed model outperforms existing video quality metrics on a data set of automotive video sequences. Furthermore, the model is generic and can be adapted to suit the needs of alternative network topologies. Future research could include the examination of alternative packet loss recovery mechanisms and their effect on perceptual quality.

Chapter 6

Conclusions and Future Work

6.1 Project Summary and Conclusions

This thesis has considered some of the issues around image and video quality assessment in the automotive environment, including methodologies for subjective and objective assessment of quality. Studies on the influence of image artifacts on both a human viewer's perception of quality and the performance of pedestrian detection algorithms have been carried out. An immediate conclusion of this thesis is that images optimized for human perception of quality are not necessarily optimal for machine vision performance. As a result, existing perceptual quality metrics do not always accurately predict the performance of detection algorithms on degraded images.

A novel IQA metric that accurately predicts the performance of pedestrian detection algorithms under varying image quality has been proposed in Chapter 3. The proposed metric operates by comparing the underlying HOG vectors in both the reference and degraded frames. The metric, termed HMSE, is computationally inexpensive and achieves higher correlation with algorithm performance for a number of recently-proposed pedestrian detection algorithms, compared to image quality algorithms such as SSIM, FSIM and IWSSIM.

The metric is full reference and hence can be used at the system design stage in order to optimize video quality for detection performance. However, such a metric is unsuitable for real-time automotive implementations since typically a reference frame is unavailable in such an environment. An analysis of image quality impairments and their relationship to the performance of pedestrian detection algorithms has been carried out. It has been shown for example that for particular levels of JPEG compressed images, a small improvement in quality can lead to a significant increase in the performance of a pedestrian detection algorithm. This is an important result, since in most automotive vision systems, the effect of such quality degradations can be mitigated by controlling the compression ratio. The results of this analysis have shown that detection performance can be optimized on distorted images by training a classifier specifically for distortion level and type. These results have been used along with spatial NSS features to develop a no reference framework for multi-classifier pedestrian detection in Chapter 4. The multi-classifier approach offers higher detection accuracy on distorted images than single classifier performance, including when evaluated on a database with real-world impairments. The proposed framework represents an improvement on approaches based on single classifiers as used in most of the existing literature; for example, a reduction in Log-Average-Miss-Rate (LAMR) from 21.82 to 17.51 was obtained compared to a classifier trained only on high quality images from the ChangeDetection data set [127]. Furthermore, since the approach operates directly on multi-scale spatial pixel data it is computationally efficient and suitable for real-time implementation.

Analysis of the perceived quality of automotive images and video sequences has been carried out through subjective quality tests in Chapter 5. The results demonstrate that saliency plays an important role in the perception of quality. Moreover, in the automotive environment, saliency is more strongly influenced by top-down, rather than bottom-up, visual processing. For example, in the subjective video test concerning packet loss impairments, pedestrians were almost always the most salient feature in a frame, regardless of the location or rate of packet losses.

A method for evaluating the perceptual blur in automotive images that have been converted from fish-eye to rectilinear images has been presented. The algorithm makes use of saliency information by utilizing a graph-based visual saliency model to weight local sharpness scores.

Finally, based on the ground truth saliency data collected from human observers, a no-reference model for evaluating the perceptual quality of video corrupted with packet losses has been proposed. Results from subjective experiments demonstrate that the proposed model correlates closer to ground truth subjective quality opinions than existing image quality metrics on a data set of automotive video sequences, achieving a Pearson Linear Correlation Coefficient (PLCC) of 0.8211 with all video sequences, compared with a PLCC of 0.4240 for the next best performing metric tested. The algorithm is computationally inexpensive and could be used in real-time automotive vision systems to monitor and adapt video quality.

6.2 Primary Contributions

The primary contributions of this thesis are listed below.

1. An evaluation of the effect of image impairments on pedestrian detection algorithm performance has been carried out. The degree to which typical automotive video quality degradations such as compression artifacts and sensor noise affect the performance of pedestrian detection algorithms has been evaluated.

2. A full reference IQA algorithm based on HOG vectors has been proposed, which correlates closely with pedestrian detection performance on degraded video frames. The metric can be used to optimize automotive video settings while mitigating the need for testing on large annotated databases.

3. A framework for no reference distortion classification has been presented. The classification metric is based on natural image statistics. In tests conducted on a large image database the algorithm correctly classified over 99% of distorted images. The classification technique has been combined with a multi-classifier approach to pedestrian detection in order to increase detection performance on degraded images.

4. A new approach for predicting the QoE of fish-eye to rectilinear transformed images has been proposed. The algorithm incorporates a model of visual saliency in order to improve correlation with human opinion scores.

5. An automotive specific video quality database has been presented consisting of 50 video sequences with associated human saliency data and MOSes. The

influence of packet loss on visual QoE for high bandwidth automotive networks has been evaluated.

6. A saliency based framework for no reference video quality assessment of packet loss degraded video has been proposed. The metric is computationally efficient and correlates well with human opinion.

6.3 Suggestions for Future Work

There are several potential areas of research that could be explored in future work.

Improving the robustness of detection algorithms to noise and compression artifacts is a pre-processing option that could be considered. In particular, it would be of interest to analyze the effect of pre-processing images with de-noising and de-blocking algorithms before extracting features for detection. From a quality perspective, such processing may improve the perceptual quality of the resulting video and could also potentially improve the detection performance of pedestrian detection algorithms. However in real-time automotive systems, implementation of these pre-processing steps would likely be challenging. It could also be of interest to evaluate reference and test HOG vectors at multiple scales to examine whether or not there is an optimal scale for image quality evaluation.

The framework for distortion categorization could be expanded to include other types of distortion such as blurring. The addition of more distortions would increase algorithm complexity and also make correct classification more difficult. Any resulting improvement in detection performance would need to be evaluated in the context of additional complexity and computation time.

The interaction between spatial scene statistics and machine vision performance is an interesting area of research which remains relatively unexplored. Future work could examine the scene statistics of automotive images in more detail. For example, typical automotive images use High Dynamic Range (HDR) image sensors with super-wide fields of view. It is known that the statistics of HDR images differ from those of low dynamic range images [161]. The super-wide field of view lenses also exhibit severe radial distortion which is likely to further alter scene statistics. The question as to whether or not distortions such as AWGN and degradations introduced by block-DCT based image compression techniques introduce the same characteristic deviations in the statistics of HDR and radially distorted images as are found in natural images is a topic that could be explored in more detail. Indeed there are hardware solutions to the wide FOV problem that could also be examined as a potential solution for automotive applications.

Evaluating both perceptual image quality and machine vision performance in sub optimal weather conditions is a research topic that is under-researched. Optimizing detection performance in heavy rain or fog, for example, would be desirable.

Finally, the development of efficient, embedded implementations of the algorithms proposed in this thesis, with real-time performance, is a further topic of interest.

Bibliography

[1] *Towards a European road safety area: policy orientations on road safety 2011-2020*, European Commission Std. [Online]. Available: http://ec.europa.eu/transport/road_safety/pdf/road_safety_citizen/road_safety_citizen_100924_en.pdf

[2] *Statistics of Road Traffic Accidents in Europe and North America*, United Nations Economic Commission for Europe Std. [Online]. Available: http://www.unece.org/trans/main/wp6/publications/stats_accidents2011.html

[3] S. Tuohy, M. Glavin, E. Jones, M. Trivedi, and L. Kilmartin, "Next generation wired intra-vehicle networks, a review," in *IEEE Intelligent Vehicles Symposium (IV), 2013*, June 2013, pp. 777–782.

[4] S. Sivaraman and M. Trivedi, "Integrated lane and vehicle detection, localization, and tracking: A synergistic approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 2, pp. 906–917, 2013.

[5] European Parliament and Council, "Directive 2010/40/EU of the European Parliament and of the Council of 7 July 2010 on the framework for the deployment of Intelligent Transport Systems in the field of road transport and for interfaces with other modes of transport Text with EEA relevance," *Official Journal of the European Union*, 2010.

[6] T. Gandhi and M. Trivedi, "Pedestrian protection systems: Issues, survey, and challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, no. 3, pp. 413 –430, Sept. 2007.

[7] X.-B. Cao, H. Qiao, and J. Keane, "A low-cost pedestrian-detection system with a single optical camera," *IEEE Transactions on Intelligent Transportation Systems*, vol. 9, no. 1, pp. 58 –67, March 2008.

[8] P. Kelly, N. E. O'Connor, and A. F. Smeaton, "Robust pedestrian detection and tracking in crowded scenes," *Image and Vision Computing*, vol. 27, no. 10, pp. 1445–1458, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0262885608000863

[9] J. Zhang and S. Gong, "People detection in low-resolution video with non-stationary background," *Image and Vision Computing*, vol. 27, no. 4, pp. 437 – 443, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0262885608001443

[10] Y. Motai, S. K. Jha, and D. Kruse, "Human tracking from a mobile agent: Optical flow and kalman filter arbitration," *Signal Processing: Image Communication*, vol. 27, no. 1, pp. 83 – 95, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0923596511000713

[11] J. Forster, X. Jiang, A. Terzis, and A. Rothermel, "Evaluation of compression algorithms for automotive stereo matching," in *IEEE Intelligent Vehicles Symposium (IV)*, June 2012, pp. 1017–1022.

[12] S. Winkler and S. Susstrunk, "Visibility of noise in natural images," in *SPIE Electronic Imaging 2004: Human Vision and Electronic Imaging IX*, vol. 5292, Jan. 2004, pp. 121–129. [Online]. Available: http://dx.doi.org/10.1117/12.526752

[13] C. Hughes, R. McFeely, P. Denny, M. Glavin, and E. Jones, "Equidistant fish-eye perspective with application in distortion centre estimation," *Image and Vision Computing*, vol. 28, no. 3, pp. 538 – 551, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0262885609001863

[14] *ITU-R Recommendation BT.500-11 Methodology for the subjective assessment of the quality of television pictures*, International Telecommunications Union Std.

[15] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[16] P. Le Callet and F. Autrusseau, "Subjective quality assessment irccyn/ivc database," 2005, http://www.irccyn.ec-nantes.fr/ivcdb/.

[17] N. Ponomarenko, V. Lukin, K. Egiazarian, J. Astola, M. Carli, and F. Battisti, "Color image database for evaluation of image quality metrics," in *IEEE 10th Workshop on Multimedia Signal Processing*, Oct. 2008, pp. 403–408.

[18] *ITU-T Recommendation P.910 Subjective video quality assessment methods for multimedia applications*, International Telecommunications Union Std., 1999.

[19] M. Leszczuk, L. Janowski, P. Romaniak, A. Glowacz, and R. Mirek, "Quality assessment for a licence plate recognition task based on a video streamed in limited networking conditions," in *Multimedia Communications, Services and*

Security, ser. Communications in Computer and Information Science, vol. 149. Springer Berlin Heidelberg, 2011, pp. 10–18.

[20] M. Leszczuk, A. Koń, J. Dumke, and L. Janowski, "Redefining ITU-T P.912 recommendation requirements for subjects of quality assessments in recognition tasks," in *Multimedia Communications, Services and Security*, ser. Communications in Computer and Information Science, vol. 287. Springer Berlin Heidelberg, 2012, pp. 188–199.

[21] D. Hertel and E. Chang, "Image quality standards in automotive vision applications," in *IEEE Intelligent Vehicles Symposium*, 2007, pp. 404–409.

[22] C. Hughes, M. Glavin, E. Jones, and P. Denny, "Wide-angle camera technology for automotive applications: a review," *IET Intelligent Transport Systems*, vol. 3, no. 1, pp. 19 –31, march 2009.

[23] A. Winterlich, V. Zlokolica, P. Denny, L. Kilmartin, M. Glavin, and E. Jones, "A saliency weighted no-reference perceptual blur metric for the automotive environment," in *Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, 2013, pp. 206–211.

[24] J. E. A. Jr. and B. Pillman, "Digital camera image formation: Introduction and hardware." Springer New York, 2013, pp. 3–44.

[25] I. Andorko, P. Corcoran, and P. Bigioi, "A dual image processing pipeline camera with ce applications," in *Consumer Electronics (ICCE), 2011 IEEE International Conference on*, Jan 2011, pp. 737–738.

[26] A. Winterlich, P. Denny, L. Kilmartin, M. Glavin, and E. Jones, "Performance optimization for pedestrian detection on degraded video using natural scene statistics," *Journal of Electronic Imaging*, vol. 23, no. 6, pp. 061 114–061 114, 2014.

[27] A. Winterlich, C. Hughes, L. Kilmartin, M. Glavin, and E. Jones, "An oriented gradient based image quality metric for pedestrian detection performance evaluation," *Signal Processing: Image Communication*, vol. 31, pp. 61–75, 2015.

[28] Z. Wang and A. Bovik, "Mean squared error: love it or leave it? A new look at signal fidelity measures," *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98–117, 2009.

[29] Z. Wang, A. Bovik, and L. Lu, "Why is image quality assessment so difficult?" in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, April 2002, p. IV.

[30] A. Beghdadi, M.-C. Larabi, A. Bouzerdoum, and K. Iftekharuddin, “A survey of perceptual image processing methods,” *Signal Processing: Image Communication*, vol. 28, no. 8, pp. 811 – 831, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0923596513000945

[31] D. Chandler and S. Hemami, “VSNR: A Wavelet-Based Visual Signal-to-Noise Ratio for Natural Images,” *IEEE Transactions on Image Processing*, vol. 16, no. 9, pp. 2284 –2298, Sept. 2007.

[32] H. Sheikh, Z. Wang, A. Bovik, and L. Cormack. Live image quality assessment database release 2. http://live.ece.utexas.edu/research/quality.

[33] N. Damera-Venkata, T. Kite, W. Geisler, B. Evans, and A. Bovik, “Image quality assessment based on a degradation model,” *IEEE Transactions on Image Processing*, vol. 9, no. 4, pp. 636 –650, Apr. 2000.

[34] R. De Freitas Zampolo and R. Seara, “A comparison of image quality metric performances under practical conditions,” in *IEEE International Conference on Image Processing (ICIP)*, vol. 3, Sept. 2005, pp. III – 1192–5.

[35] E. Peli, “Contrast in complex images,” *Journal of the Optical Society of America A*, vol. 7, no. 10, pp. 2032–2040, Oct 1990.

[36] H. Sheikh, M. Sabir, and A. Bovik, “A statistical evaluation of recent full reference image quality assessment algorithms,” *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440 –3451, Nov. 2006.

[37] Z. Wang, L. Lu, and A. Bovik, “Video quality assessment based on structural distortion measurement,” *Signal Processing:Image Communication*, vol. 19, no. 2, pp. 121–132, 2004.

[38] H. R. Sheikh, A. C. Bovik, and G. de Veciana, “An information fidelity criterion for image quality assessment using natural scene statistics,” *IEEE transactions on Image Processing*, vol. 14, no. 12, pp. 2117–2128, 2005.

[39] H. Sheikh and A. Bovik, “Image information and visual quality,” *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430 –444, Feb. 2006.

[40] H. Zhang, Y. Huang, X. Chen, and D. Deng, “MLSIM: A multi-level similarity index for image quality assessment,” *Signal Processing:Image Communication*, vol. 28, no. 10, pp. 1464 – 1477, 2013.

[41] J. Prewitt, *Object enhancement and extraction in picture Processing and Psychopictorics.* Academic Press, New York, 1970.

[42] D. B. Russakoff, C. Tomasi, T. Rohlfing, and C. R. Maurer Jr, "Image similarity using mutual information of regions," in *Computer Vision-ECCV 2004*. Springer, 2004, pp. 596–607.

[43] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "A comprehensive evaluation of full reference image quality assessment algorithms," in *19th IEEE International Conference on Image Processing (ICIP)*, Sept. 2012, pp. 1477–1480.

[44] L. Zhang and H. Li, "SR-SIM: A fast and high performance IQA index based on spectral residual," in *Image Processing (ICIP), 2012 19th IEEE International Conference on*, 2012, pp. 1473–1476.

[45] L. Zhang, D. Zhang, X. Mou, and D. Zhang, "FSIM: A Feature Similarity Index for Image Quality Assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[46] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 5, pp. 1185–1198, 2011.

[47] Q. Li and Z. Wang, "Reduced-reference image quality assessment using divisive normalization-based image representation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 202–211, 2009.

[48] I. Gunawan and M. Ghanbari, "Reduced-reference video quality assessment using discriminative local harmonic strength with motion consideration," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 1, pp. 71–83, 2008.

[49] Video Quality Experts Group. (2000) FRTV test Phase 1 Video Sequences. VQEG. [Online]. Available: http://www.vqeg.org/

[50] S. Wolf and M. H. Pinson, "Low bandwidth reduced reference video quality monitoring system," in *First International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 2005, pp. 23–25.

[51] S. Wolf and M. Pinson, "Video Quality Measurement Techniques, NTIA Report 02-392," NTIA, Tech. Rep., 2002.

[52] M. Carnec, P. Le Callet, and D. Barba, "An image quality assessment method based on perception of structural information," in *International Conference on Image Processing (ICIP)*, vol. 3. IEEE, 2003, pp. III–185.

[53] M. Carnec, P. Le Callet, and D. Barba, "Visual features for image quality assessment with reduced reference," in *International Conference on Image Processing (ICIP).*, vol. 1. IEEE, 2005, pp. I–421.

[54] Z. Wang and E. P. Simoncelli, “Reduced-reference image quality assessment using a wavelet-domain natural image statistic model,” in *Proc. of SPIE Human Vision and Electronic Imaging*, 2005, pp. 149–159.

[55] M. J. Wainwright and E. P. Simoncelli, “Scale mixtures of gaussians and the statistics of natural images.” in *Advances in Neural Information Processing Systems*. Citeseer, 1999, pp. 865–861.

[56] R. Soundararajan and A. Bovik, “Video quality assessment by reduced reference spatio-temporal entropic differencing,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 4, pp. 684–694, 2013.

[57] R. Soundararajan and A. Bovik, “RRED Indices: Reduced Reference Entropic Differencing for Image Quality Assessment,” *IEEE Transactions on Image Processing*, vol. 21, no. 2, pp. 517–526, 2012.

[58] L. Ma, S. Li, and K. N. Ngan, “Reduced-reference video quality assessment of compressed video sequences,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 10, pp. 1441–1456, 2012.

[59] H. Liu, R. Zunino, I. Heynderickx, J. Redi, and H. Alers, “Efficient neural-network based no-reference approach to an overall quality metric for jpeg and jpeg2000 compressed images,” *Journal of Electronic Imaging*, vol. 20, no. 4, pp. 043 007–043 007–15, 2011.

[60] S. Liu and A. Bovik, “Efficient dct-domain blind measurement and reduction of blocking artifacts,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 12, pp. 1139–1149, 2002.

[61] H. Liu, N. Klomp, and I. Heynderickx, “A no-reference metric for perceived ringing artifacts in images,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 4, pp. 529–539, 2010.

[62] P. Romaniak, L. Janowski, M. Leszczuk, and Z. Papir, “Perceptual quality assessment for H.264/AVC compression,” in *Consumer Communications and Networking Conference (CCNC)*, 2012, pp. 597–602.

[63] Z. Wang, H. Sheikh, and A. Bovik, “No-reference perceptual quality assessment of JPEG compressed images,” in *IEEE International Conference on Image Processing*, vol. 1, 2002, pp. 477–480.

[64] E. Ong, W. Lin, Z. Lu, S. Yao, X. Yang, and L. Jiang, “No-reference JPEG-2000 image quality metric,” in *International Conference on Multimedia and Expo*, vol. 1. IEEE, 2003, pp. I–545.

[65] H. Sheikh, A. Bovik, and L. Cormack, "No-reference quality assessment using natural scene statistics: JPEG2000," *IEEE Transactions on Image Processing*, vol. 14, no. 11, pp. 1918 –1927, Nov. 2005.

[66] X. Zhu and P. Milanfar, "A no-reference sharpness metric sensitive to blur and noise," in *International Workshop on Quality of Multimedia Experience (QoMEX)*, July 2009, pp. 64 –69.

[67] M.-J. Chen and A. Bovik, "No-reference image blur assessment using multiscale gradient," in *International Workshop on Quality of Multimedia Experience(QoMEX)*, July 2009, pp. 70 –74.

[68] N. Sadaka, L. Karam, R. Ferzli, and G. Abousleman, "A no-reference perceptual image sharpness metric based on saliency-weighted foveal pooling," in *15th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2008, pp. 369–372.

[69] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "A no-reference perceptual blur metric," in *International Conference on Image Processing*, vol. 3, 2002, pp. III–57–III–60 vol.3.

[70] G. Valenzise, S. Magni, M. Tagliasacchi, and S. Tubaro, "No-reference pixel video quality monitoring of channel-induced distortion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 4, pp. 605–618, 2012.

[71] F. Yang, S. Wan, Q. Xie, and H. R. Wu, "No-reference quality assessment for networked video via primary analysis of bit stream," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 11, pp. 1544–1554, 2010.

[72] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "Perceptual blur and ringing metrics: application to JPEG2000," *Signal Processing: Image Communication*, vol. 19, no. 2, pp. 163–172, 2004.

[73] E. Allen, S. Triantaphillidou, and R. Jacobson, "Perceptibility and acceptability of JPEG 2000 compressed images of various scene types," in *Image Quality and System Performance XI*, vol. 9016, 2014, pp. 90 160W–90 160W–15. [Online]. Available: http://dx.doi.org/10.1117/12.2042582

[74] A. Hyvärinen, J. Hurri, and P. O. Hoyer, *Natural Image Statistics.* Springer, 2009.

[75] A. Mittal, A. Moorthy, and A. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[76] A. Moorthy and A. Bovik, "Statistics of natural image distortions," in *2010 IEEE International Conference on Acoustics Speech and Signal Processing*, March 2010, pp. 962–965.

[77] A. Moorthy and A. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, Dec 2011.

[78] M. Saad, A. Bovik, and C. Charrier, "A DCT statistics-based blind image quality index," *IEEE Signal Processing Letters*, vol. 17, no. 6, pp. 583 –586, June 2010.

[79] M. Saad, A. Bovik, and C. Charrier, "DCT statistics model-based blind image quality assessment," in *18th IEEE International Conference on Image Processing*, Sept. 2011, pp. 3093–3096.

[80] P. Ye and D. Doermann, "No-reference image quality assessment using visual codebooks," *IEEE Transactions on Image Processing*, vol. 21, no. 7, pp. 3129–3138, July 2012.

[81] A. Mittal, R. Soundararajan, and A. Bovik, "Making a completely blind image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013.

[82] Y. Zhang and D. M. Chandler, "No-reference image quality assessment based on log-derivative statistics of natural scenes," *Journal of Electronic Imaging*, vol. 22, no. 4, 2013.

[83] I. Alonso, D. Llorca, M. Sotelo, L. Bergasa, P. R. de Toro, J. Nuevo, M. Ocana, and M. Garrido, "Combination of feature extraction methods for SVM pedestrian detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, no. 2, pp. 292 –307, June 2007.

[84] R. Benenson, M. Mathias, T. Tuytelaars, and L. Van Gool, "Seeking the strongest rigid detector," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013, pp. 3666–3673.

[85] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 886–893.

[86] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral Channel Features," in *British Machine Vision Conference*, 2009.

[87] P. Dollar, C. Wojek, B. Schiele, and P. Perona, “Pedestrian detection: An evaluation of the state of the art,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743–761, 2012.

[88] R. P. Loce, E. A. Bernal, W. Wu, and R. Bala, “Computer vision in roadway transportation systems: a survey,” *Journal of Electronic Imaging*, vol. 22, no. 4, p. 041121, 2013.

[89] S. Gunasekar, J. Ghosh, and A. C. Bovik, “Face detection on distorted images using perceptual quality-aware features,” in *Human Vision and Electronic Imaging XIX*, vol. 9014, 2014, pp. 90 141E–90 141E–13. [Online]. Available: http://dx.doi.org/10.1117/12.2037343

[90] A. Tsifouti, S. Triantaphillidou, E. Bilissi, and M.-C. Larabi, “Acceptable bit-rates for human face identification from CCTV imagery,” vol. 8653, 2013. [Online]. Available: http://dx.doi.org/10.1117/12.2004140

[91] T. Hase, W. Hintermaier, A. Frey, T. Strobel, U. Baumgarten, and E. Steinbach, “Influence of image/video compression on night vision based pedestrian detection in an automotive application,” in *IEEE 73rd Vehicular Technology Conference (VTC Spring)*, May 2011, pp. 1–5.

[92] D. Gavrila and V. Philomin, “Real-time object detection for smart vehicles,” in *Proceedings of the Seventh IEEE International Conference on Computer Vision, 1999.*, vol. 1, 1999, pp. 87–93 vol.1.

[93] B. Wu and R. Nevatia, “Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors,” in *Tenth IEEE International Conference on Computer Vision (ICCV)*, vol. 1, Oct. 2005, pp. 90–97 Vol. 1.

[94] P. Sabzmeydani and G. Mori, “Detecting pedestrians by learning shapelet features,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2007, pp. 1–8.

[95] C. Wojek and B. Schiele, “A performance evaluation of single and multi-feature people detection,” in *Pattern Recognition (DAGM)*, May 2008.

[96] G. Mori, S. Belongie, and J. Malik, “Efficient shape matching using shape contexts,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 11, pp. 1832–1837, Nov. 2005.

[97] X. Wang, T. Han, and S. Yan, “An HOG-LBP human detector with partial occlusion handling,” in *IEEE 12th International Conference on Computer Vision*, Sept. 2009, pp. 32–39.

[98] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, July 2002.

[99] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[100] X. Yong, L. Zhang, Z. Song, Y. Hu, L. Zheng, and J. Zhang, "Real-time vehicle detection based on haar features and pairwise geometrical histograms," in *IEEE International Conference on Information and Automation (ICIA)*, June 2011, pp. 390–395.

[101] W. Yao and Z. Deng, "A robust pedestrian detection approach based on shapelet feature and haar detector ensembles," *Tsinghua Science and Technology*, vol. 17, no. 1, pp. 40–50, Feb. 2012.

[102] Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, 2006, pp. 1491–1498.

[103] F. Porikli, "Integral histogram: A fast way to extract histograms in cartesian spaces," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Washington, DC, USA: IEEE Computer Society, 2005, pp. 829–836.

[104] P. Dollar, S. Belongie, and P. Perona, "The fastest pedestrian detector in the west," in *Proceedings of the British Machine Vision Conference*. BMVA Press, 2010, pp. 68.1–68.11, doi:10.5244/C.24.68.

[105] G. Xu, X. Wu, L. Liu, and Z. Wu, "Real-time pedestrian detection based on edge factor and histogram of oriented gradient," in *IEEE International Conference on Information and Automation (ICIA)*, 2011, pp. 384–389.

[106] A. Chavan and S. Yogamani, "Real-time DSP implementation of pedestrian detection algorithm using HOG features," in *12th International Conference on ITS Telecommunications (ITST)*, 2012, pp. 352–355.

[107] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[108] M. Hussein, F. Porikli, and L. Davis, "A comprehensive evaluation framework and a comparative study for human detectors," *IEEE Transactions on Intelligent Transportation Systems*, vol. 10, no. 3, pp. 417 –427, Sept. 2009.

[109] A. Ruta, Y. Li, and X. Liu, "Robust class similarity measure for traffic sign recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 11, no. 4, pp. 846 –855, Dec. 2010.

[110] R. Benenson, M. Mathias, R. Timofte, and L. Van Gool, "Pedestrian detection at 100 frames per second," in *IEEE Conferenceon Computer Vision and Pattern Recognition*, June 2012, pp. 2903–2910.

[111] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[112] P. Dollár, "Piotr's Image and Video Matlab Toolbox (PMT)," https://github.com/pdollar/toolbox.

[113] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, pp. 2179–2195, 2009.

[114] J. Marin, D. Vazquez, D. Geronimo, and A. Lopez, "Learning appearance in virtual scenarios for pedestrian detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2010, pp. 137 –144.

[115] L. Wang, J. Shi, G. Song, and I.-F. Shen, "Object detection combining recognition and segmentation," in *Proceedings of the 8th Asian conference on Computer vision*, ser. ACCV'07, 2007, pp. 189–199.

[116] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, "Tid2008-a database for evaluation of full-reference visual quality assessment metrics," *Advances of Modern Radioelectronics*, vol. 10, no. 4, pp. 30–45, 2009.

[117] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, "Live image quality assessment database release 2," 2005.

[118] D. G. Lee, S. M. Jung, and M.-S. Lim, "System on chip design of embedded controller for car black box," in *IEEE Intelligent Vehicles Symposium*, 2007, pp. 1174–1177.

[119] N. C. Francisco, N. M. Rodrigues, E. A. da Silva, and S. M. de Faria, "A generic post-deblocking filter for block based image compression algorithms," *Signal Processing: Image Communication*, vol. 27, no. 9, pp. 985 – 997, 2012.

[120] A. B. Watson, J. A. Solomon, A. J. Ahumada, Jr., and A. Gale, "Discrete cosine transform (DCT) basis function visibility: effects of viewing distance

and contrast masking," vol. 2179, 1994, pp. 99–108. [Online]. Available: http://dx.doi.org/10.1117/12.172661

[121] G. Wallace, "The JPEG still picture compression standard," *IEEE Transactions on Consumer Electronics*, vol. 38, no. 1, pp. xviii–xxxiv, Feb. 1992.

[122] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval.* New York, NY, USA: McGraw-Hill, Inc., 1986.

[123] R. Wagner, M. Gabb, J. Forster, R. Schweiger, and A. Rothermel, "Improving detector performance by learning from compressed samples," in *IEEE International Conference on Consumer Electronics*, 2012, pp. 200–204.

[124] R. A. Fisher, "On the probable error of a coefficient of correlation deduced from a small sample," *Metron*, vol. 1, pp. 3–32, 1921.

[125] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1532–1545, 2014.

[126] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2009, pp. 304 –311.

[127] N. Goyette, P. Jodoin, F. Porikli, J. Konrad, and P. Ishwar, "Changedetection.net: A new change detection benchmark dataset," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2012, pp. 1–8.

[128] M. Schöberl, A. Brückner, S. Foessel, and A. Kaup, "Photometric limits for digital camera systems," *Journal of Electronic Imaging*, vol. 21, no. 2, pp. 020 501–1–020 501–3, 2012.

[129] D. L. Ruderman and W. Bialek, "Statistics of natural images: Scaling in the woods," *Phys. Rev. Lett.*, vol. 73, pp. 814–817, Aug 1994. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevLett.73.814

[130] A. Moorthy and A. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Processing Letters*, vol. 17, no. 5, pp. 513–516, 2010.

[131] A. Srivastava, A. B. Lee, E. P. Simoncelli, and S.-C. Zhu, "On advances in statistical modeling of natural images," *Journal of Mathematical Imaging and Vision*, vol. 18, pp. 17–33, 2003.

[132] M. Leszczuk, I. Stange, and C. Ford, “Determining image quality requirements for recognition tasks in generalized public safety video applications: Definitions, testing, standardization, and current trends,” in *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, June 2011, pp. 1–5.

[133] A. Borji, D. N. Sihite, and L. Itti, “Quantitative analysis of human-model agreement in visual saliency modeling: a comparative study,” *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 55–69, 2013.

[134] *Design Standards and Guidelines*, Bureau of street lighting Std.

[135] S. E. Susstrunk and S. Winkler, “Color image quality on the internet,” in *SPIE Electronic Imaging 2004: Internet Imaging V*, vol. 5304, Jan. 2003, pp. 118–131. [Online]. Available: http://dx.doi.org/10.1117/12.537804

[136] D. Scaramuzza, A. Martinelli, and R. Siegwart, “A flexible technique for accurate omnidirectional camera calibration and structure from motion,” in *IEEE International Conference on Computer Vision Systems (ICVS)*, Jan. 2006.

[137] H. de Ridder and S. Endrikhovski, “Image quality is fun: Reflections on fidelity, usefulness and naturalness,” *SID Symposium Digest of Technical Papers*, vol. 33, no. 1, pp. 986–989, 2002.

[138] N. Narvekar and L. Karam, “A no-reference perceptual image sharpness metric based on a cumulative probability of blur detection,” in *International Workshop on Quality of Multimedia Experience (QoMEX)*, July, pp. 87–91.

[139] X. Feng, T. Liu, D. Yang, and Y. Wang, “Saliency based objective quality assessment of decoded video affected by packet losses,” in *15th IEEE International Conference on Image Processing (ICIP)*, Oct., pp. 2560–2563.

[140] X. Hou, J. Harel, and C. Koch, “Image signature: Highlighting sparse salient regions,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 194–201, 2012.

[141] J. Harel, C. Koch, and P. Perona, “Graph-based visual saliency,” in *Advances in Neural Information Processing Systems 19*. MIT Press, 2007, pp. 545–552.

[142] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[143] L. L. Bello, “The case for ethernet in automotive communications,” *Special Interest Group on Embedded Systems Review*, vol. 8, no. 4, pp. 7–15, Dec 2011. [Online]. Available: http://doi.acm.org/10.1145/2095256.2095257

[144] T. R. Henderson, S. Roy, S. Floyd, and G. F. Riley, "ns-3 Project Goals," in *ACM Proceeding from the 2006 workshop on ns-2: the IP network simulator*, 2006.

[145] S. Tuohy, M. Glavin, C. Hughes, E. Jones, and L. Kilmartin, "An ns-3 based simulation testbed for in-vehicle communication networks," in *27th Annual UK Performance Engineering Workshop, Bradford, UK*, 2011.

[146] S. Tuohy, A. Winterlich, P. Denny, B. McGinley, M. Glavin, E. Jones, and L. Kilmartin, "Evaluating the influence of packet loss on visual quality of perception for high bandwidth automotive networks," *Signal Processing: Image Communication*, vol. to appear, 2016. [Online]. Available: 10.1016/j.image.2016.01.004

[147] H.-T. Lim, B. Krebs, L. Volker, and P. Zahrer, "Performance evaluation of the inter-domain communication in a switched Ethernet based in-car network," *IEEE 36th Conference on Local Computer Networks*, pp. 101–108, Oct. 2011.

[148] H.-T. Lim, L. Völker, and D. Herrscher, "Challenges in a future IP/Ethernet-based in-car network for real-time applications," in *48th ACM/EDAC/IEEE Design Automation Conference (DAC)*, June 2011, pp. 7–12.

[149] G. Alderisi, G. Patti, and L. Bello, "Introducing support for scheduled traffic over IEEE audio video bridging networks," in *18th IEEE International Conference on Emerging Technologies and Factory Automation*, 2013. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6647943

[150] C. Perkins, O. Hodson, and V. Hardman, "A survey of packet-loss recovery techniques for streaming audio," *IEEE Network*, vol. 12, no. 5, pp. 40–48, September 1998.

[151] N. Feamster and H. Balakrishnan, "Packet loss recovery for streaming video," in *12th International Packet Video Workshop.* PA: Pittsburgh, 2002, pp. 9–16.

[152] T. Liu, Y. Wang, J. Boyce, Z. Wu, and H. Yang, "Subjective quality evaluation of decoded video in the presence of packet losses," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 1, April 2007, pp. I–1125–I–1128.

[153] M. Leszczuk, L. Janowski, P. Romaniak, and Z. Papir, "Assessing quality of experience for high definition video streaming under diverse packet loss patterns." *Signal Processing: Image Communication*, vol. 28, no. 8, pp. 903–916, 2013.

[154] M. H. Pinson, M. Barkowsky, and P. Le Callet, "Selecting scenes for 2D and 3D subjective video quality tests," *EURASIP Journal on Image and Video Processing*, vol. 2013, no. 1, pp. 1–12, 2013.

[155] C. Fenimore, J. Libert, and S. Wolf, “Perceptual effects of noise in digital video compression,” *SMPTE Motion Imaging Journal*, vol. 109, no. 3, pp. 178–187, 2000. [Online]. Available: http://journal.smpte.org/content/109/3/178.abstract

[156] The EyeTribe, *http://dev.theeyetribe.com/dev/*.

[157] H. Liu and I. Heynderickx, “Visual attention in objective image quality assessment: Based on eye-tracking data,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 7, pp. 971–982, July 2011.

[158] O. LeMeur, A. Ninassi, P. L. Callet, and D. Barba, “Do video coding impairments disturb the visual attention deployment?” *Signal Processing: Image Communication*, vol. 25, no. 8, pp. 597 – 609, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0923596510000585

[159] J. Redi and I. Heynderickx, “Image quality and visual attention interactions: Towards a more reliable analysis in the saliency space,” in *Third International Workshop on Quality of Multimedia Experience (QoMEX)*, Sept 2011, pp. 201–206.

[160] M. Saad, A. Bovik, and C. Charrier, “Blind prediction of natural video quality,” *IEEE Transactions on Image Processing*, vol. 23, no. 3, pp. 1352–1365, March 2014.

[161] T. Pouli, D. Cunningham, and E. Reinhard, “Statistical regularities in low and high dynamic range images,” in *ACM Symposium on Applied Perception in Graphics and Visualization (APGV)*, July 2010.

Appendix A

Experimental Details

A.1 Computing the HOG feature

Matlab code to efficiently compute the HOG feature is available as part of Piotr's Image and Video Matlab toolbox referenced in chapter 3. For the experiments in this thesis the default parameters were used to extract the HOG vector, H. Namely a binsize of width 8 pixels, 9 oriented gradients (nOrients = 9), and a histogram clipping threshold of 0.2. Thus for an input image I, with dimensions $h \times w$, the size of the computed feature vector H is given by:

$$floor([\frac{h}{binSize}, \frac{w}{binSize}, nOrients \times 4]). \tag{A.1}$$

Each binSize x binSize region, computes a histogram of gradients, with each gradient quantized by its angle and weighted by its magnitude. For colour images, the gradient is computed separately for each colour channel and the one with maximum magnitude is used. Tri-linear interpolation is used to place each gradient in the appropriate spatial and orientation bin. For each resulting histogram (with nOrients bins), four different normalizations are computed using adjacent histograms, resulting in a nOrients x 4 length feature vector for each region. To compute the normalizations, first for each block of adjacent 2 x 2 histograms their L2 norm is computed. Each histogram thus has 4 different normalization values associated with it. Each histogram bin is then normalized by each of the 4 different L2 norms, resulting in a 4 times expansion of the number of bins. Finally, any bin whose value is bigger than the clipping threshold is set to the threshold value.

A.2 Algorithm Retraining

The quality parameters used for creating the data sets of distorted images in Chapter 3 are listed in table A.1. Distortion levels used for the experiments in Chapter 4 are listed in table A.2

All three pedestrian detection algorithms evaluated in this thesis utilize a form of HOG vector for pedestrian detection. The default parameters for each algorithm were used. All algorithms use a binwidth (or cell size) of 8 pixels, 9 oriented gradients, and a clipping threshold of 0.2. A spatial stride of 4 pixels was used between detection windows.

In Chapter 3 distorted versions of the INRIA training set were used to retrain each detection algorithm. For the AWGN trained model, each image in the training set was distorted with AWGN of variance of 0.002. For the JPEG trained model, all training images were JPEG compressed with Quality factor Q=50. The multi-quality model was derived from training on all three sets of the training images (reference, AWGN degraded and JPEG compressed.)

In Chapter 4, each algorithm was retrained on distorted versions of the INRIA and Caltech data sets. For each AWGN trained model, AWGN was added to every image in the training set so that it's NIQE score was within a particular range. Similarly, for every JPEG trained model, every image in the data set was compressed so that its NIQE score was within the desired range. Models were trained for all NIQE ranges in table A.3

Table A.1: Chapter 3: Quality parameters for Distorted data sets

Level	**AWGN (variance)**	**JPEG (Q)**	**JP2 (CR)**
1	0.00001	95	2
2	0.00005	90	3
3	0.0001	85	4
4	0.0002	80	5
5	0.0004	75	6
6	0.0006	70	7
7	0.0008	65	8
8	0.001	60	9
8	0.0015	55	10
10	0.002	50	12
11	0.003	46	14
12	0.004	42	16
13	0.005	38	18
14	0.006	34	20
15	0.007	30	22
16	0.008	28	24
17	0.009	26	26
18	0.01	24	28
19	0.012	22	30
20	0.014	20	34
21	0.016	19	38
22	0.02	18	42
23	0.025	17	46
24	0.03	16	50
25	0.035	15	55
26	0.04	14	60
27	0.045	13	65
28	0.05	12	70
29	0.055	11	75
30	0.06	10	80
31	0.07	9	85
32	0.075	8	90
33	0.08	7	95
34	0.085	6	100
35	0.09	5	150
36	0.095	4	200
37	0.1	3	400
38	0.2	2	600
39	0..3	1	800
40	0.5	0	1000

Table A.2: Chapter 4: Quality parameters for Distorted data sets

Level	**AWGN (variance)**	**JPEG (Q)**
1	0.00005	60
2	0.0006	50
3	0.0015	30
4	0.003	24
5	0.006	19
6	0.009	15
7	0.014	12
8	0.025	8
8	0.045	6
10	0.085	5
11	0.2	4
12	0.3	1

Table A.3: Chapter 4: Quality parameters for training models

Distortion type	**NIQE range**
Reference	0-4
AWGN	4-6
AWGN	6-8
AWGN	8-10
AWGN	10-12
AWGN	12-14
AWGN	> 14
JPEG	4-6
JPEG	6-8
JPEG	8-10
JPEG	> 10

Appendix B

Matlab Scripts and Functions

Matlab scripts to generate degraded images.

```
% File location for INRIA positive samples. This code assumes that the
% INRIA database has been stored at the particular location indicated
% below.
% This function is used to generate the JPEG compressed database used
% for the experiments reported in Chapter 3.

listing = dir('C:\INRIAPerson\Test\pos\*.png');
filepath = 'C:\INRIAPerson\test\jpeg\';
% 40 quality parameters
quality =
[95,90,85,80,75,70,65,60,55,50,46,42,38,34,30,28,26,24,22,20,19,18,17,
16,15,14,13,12,11,10,9,8,7,6,5,4,3,2,1,0];

for j = 1:40
    batch = num2str(quality(j));
    folder = strcat(filepath,batch);

% check folder exists
      if ~exist(folder, 'dir')
      mkdir(folder);
      end

% for each of the 288 positive INRIA samples
    for i = 1:288
        filename = strcat(filepath,batch,'\',listing(i).name);
        [a,b,c] = fileparts(filename);
        filename = strcat(a,'\',b,'.jpg');

 %read files sequentially
        getfile = strcat('C:\INRIAPerson\Test\pos\',listing(i).name);
        I = imread(getfile);

% Write degraded image file to folder
        imwrite(I,filename,'Quality',quality(j));
     end
end
```

```
% File location for INRIA positive samples. This code assumes that the
% INRIA database has been stored at the particular location indicated
% below.
% This function is used to generate the JPEG2K compressed database
used % for the experiments reported in Chapter 3.

% File location for INRIA positive samples
listing = dir('C:\INRIAPerson\Test\pos\*.png');
filepath = 'C:\INRIAPerson\test\jpeg2k\';

% 40 compression parameters
compression =
[2,3,4,5,6,7,8,9,10,12,14,16,18,20,22,24,26,28,30,34,38,42,46,50,55,60
,65,70,75,80,85,90,95,100,150,200,400,600,800,1000];

for j = 1:40
    batch = num2str(compression(j));
    folder = strcat(filepath,batch);

% check folder exists
    if ~exist(folder, 'dir')
    mkdir(folder);
    end

% for each of the 288 positive INRIA samples
    for i =1:288
        filename = strcat(filepath,batch,'\',listing(i).name);
        [a,b,c] = fileparts(filename);
        filename = strcat(a,'\',b,'.jp2');

%read files sequentially
        getfile = strcat('C:\INRIAPerson\Test\pos\',listing(i).name);
        I = imread(getfile);

% Write degraded image file to folder
        imwrite(I,filename,'compression',compression(j));

     end
end
```

```
% File location for INRIA positive samples. This code assumes that the
% INRIA database has been stored at the particular location indicated
% below.
% This function is used to generate the noise database used for the
% experiments reported in Chapter 3.

% File location for INRIA positive samples
listing = dir('C:\INRIAPerson\Test\pos\*.png');
filepath = 'C:\INRIAPerson\test\noise\';

% 40 quality (variance) parameters
for j = 1:40
    variance =
[0.00001,0.00005,0.0001,0.0002,0.0004,0.0006,0.0008,0.0010,0.0015...

0.002,0.003,0.004,0.005,0.006,0.007,0.008,0.009,0.01,0.012,0.014...

0.016,0.02,0.025,0.03,0.035,0.04,0.045,0.05,0.055,0.06,0.07,0.075,...
    0.08,0.085,0.09,0.095,0.1,0.2,0.3,0.5];

    batch = num2str(j);
    folder = strcat(filepath,batch);

% check folder exists
    if ~exist(folder, 'dir')
    mkdir(folder);
    end

% for each of the 288 positive INRIA samples
    for i = 1:288
        filename = strcat(filepath,batch,'\',listing(i).name);
        [a,b,c] = fileparts(filename);
        filename = strcat(a,'\',b,'.bmp');

%read files sequentially
        getfile = strcat('C:\INRIAPerson\Test\pos\',listing(i).name);
        I = imread(getfile);
        I = imnoise(I,'gaussian',0,variance(j));

% Write degraded image file to folder
        imwrite(I,filename);
    end
end
```

```
% Routine to evaluate detections against ground truth. Piotr Dollar’s
% image processing toolbox is required and is available to download
% at: https://github.com/pdollar/toolbox
%This code was used for the experiments reported in Chapter 4

% preload all classifiers (detection)
load models\AcfInria1000Detector.mat;
load models\AcfInriaNoise6to8Detector.mat;
load models\AcfInriaJPEG6to8Detector.mat;
load models\AcfInriaNoise8to10Detector.mat;
load models\AcfInriaJPEG8to10Detector.mat;
load models\AcfInriaNoise10to12Detector.mat;
load models\AcfInriaJPEG10plusDetector.mat;
load models\AcfInriaNoise12to14Detector.mat;
load models\AcfInriaNoise14plusDetector.mat;
load models\AcfInriaOptimalDetector.mat;
load AcfInriaRobustDetector.mat;
load AcfInriaRobustPristinePosDetector.mat');

datadir =
'C:\Pedestrian_detection\cwd2014\pedestriandetectiondataset\pedestrian
s\input\';
d =
dir('C:\Pedestrian_detection\cwd2014\pedestriandetectiondataset\pedest
rians\input\*.jpg');
[n,~] = size(d)
writeFile = 'MultiClassifierNew.txt';

%ground truth directory
gtDir =
'C:\Pedestrian_detection\cwd2014\pedestriandetectiondataset\pedestrian
s\gt';

% set parameters
    blocksizerow    = 96;
    blocksizecol    = 96;
    blockrowoverlap = 0;
    blockcoloverlap = 0;
    thr = 0.5;mul =0; ref = 10.^(-2:.25:0);
    lims = [3.1e-3 1e1 .05 1];

display('detecting pedestrians...');

% display progress...
 for i = 1:n
     val = mod(i,100);
     if (val == 0)
         fprintf([num2str(i) '/' num2str(n) ' images processed\n']);
     end
 I = imread([datadir d(i).name]);
```

```
%load pristine model

load C:\INRIAPerson\niqe_release\modelparameters.mat;
quality =
computequality(I,blocksizerow,blocksizecol,blockrowoverlap,blockcolove
rlap,mu_prisparam,cov_prisparam);

%select appropriate classifier
 if (quality <= 6)
  detector = PristineDetector;

 elseif (quality <=8)
    load C:\INRIAPerson\niqe_release\modelparameters_jpeg6to8.mat;
    jpeg =
computequality(I,blocksizerow,blocksizecol,blockrowoverlap,blockcolove
rlap,mu_prisparam,cov_prisparam);
    load C:\INRIAPerson\niqe_release\modelparameters_noise6to8.mat;
    noise =
computequality(I,blocksizerow,blocksizecol,blockrowoverlap,blockcolove
rlap,mu_prisparam,cov_prisparam);
          if (noise<=jpeg)
             detector = Noise6to8Detector;
          else
              detector = JPEG6to8Detector;
          end

 elseif (quality <= 10)
     load C:\INRIAPerson\niqe_release\modelparameters_jpeg8to10.mat;
     jpeg =
computequality(I,blocksizerow,blocksizecol,blockrowoverlap,blockcolove
rlap,mu_prisparam,cov_prisparam);
     load C:\INRIAPerson\niqe_release\modelparameters_noise8to10.mat;
     noise =
computequality(I,blocksizerow,blocksizecol,blockrowoverlap,blockcolove
rlap,mu_prisparam,cov_prisparam);
          if (noise<=jpeg)
           detector = Noise8to10Detector;
          else
           detector = JPEG8to10Detector;
          end

 elseif (quality <= 12)
    load C:\INRIAPerson\niqe_release\modelparameters_jpeg10to12.mat
    jpeg =
computequality(I,blocksizerow,blocksizecol,blockrowoverlap,blockcolove
rlap,mu_prisparam,cov_prisparam);
    load C:\INRIAPerson\niqe_release\modelparameters_noise10to12.mat
    noise =
computequality(I,blocksizerow,blocksizecol,blockrowoverlap,blockcolove
rlap,mu_prisparam,cov_prisparam);
          if (noise<=jpeg)
```

```
              detector = Noise10to12Detector;
          else
              detector = JPEG10plusDetector;
          end

 elseif (quality <= 14)
      load C:\INRIAPerson\niqe_release\modelparameters_jpeg12to14.mat;
      jpeg =
computequality(I,blocksizerow,blocksizecol,blockrowoverlap,blockcolove
rlap,mu_prisparam,cov_prisparam);
     load C:\INRIAPerson\niqe_release\modelparameters_noise12to14.mat;
      noise =
computequality(I,blocksizerow,blocksizecol,blockrowoverlap,blockcolove
rlap,mu_prisparam,cov_prisparam);
          if (noise<=jpeg)
             detector = Noise12to14Detector;
          else
             detector = JPEG10plusDetector;
          end

 else
      load C:\INRIAPerson\niqe_release\modelparameters_jpeg14plus.mat;
      jpeg =
computequality(I,blocksizerow,blocksizecol,blockrowoverlap,blockcolove
rlap,mu_prisparam,cov_prisparam);
     load C:\INRIAPerson\niqe_release\modelparameters_noise14plus.mat;
      noise =
computequality(I,blocksizerow,blocksizecol,blockrowoverlap,blockcolove
rlap,mu_prisparam,cov_prisparam);
          if (noise<=jpeg)
              detector = Noise14plusDetector;
          else
              detector = JPEG10plusDetector;

          end
end


%bounding box
 bbs = acfDetect(I, detector);
 [m,~] = size(bbs);
 x = zeros(m,1);
 x(:,1) = i; % i is image id, to be pre-pended to bbs
 dtI = cat(2,x,bbs);

%write results to file
dlmwrite(writeFile,dtI,'-append');

end
```

```
display('Evaluating detection performance...');
% run evaluation using bbGt
[gt,dt] = bbGt('loadAll',gtDir,writeFile);
[gt,dt] = bbGt('evalRes',gt,dt,thr,mul);
[fp,tp,score,miss] = bbGt('compRoc',gt,dt,1,ref);
miss=exp(mean(log(max(1e-10,1-miss)))); roc=[score fp tp];

%plot roc
figure(1);hold on; plotRoc([fp tp],'logx',1,'logy',1,'xLbl','fppi',...
  'lims',lims,'color','g','smooth',1,'fpTarget',ref);
title(sprintf('log-average miss rate = %.2f%%',miss*100));
```

```
% The function (written by Zhou Wang) used to evaluate the SSIM index
% for the experiments reported in Chapter 3 is reproduced here for
% convenience.

function [mssim, ssim_map] = ssim_index(img1, img2, K, window, L)

%======================================================================
%SSIM Index, Version 1.0
%Copyright(c) 2003 Zhou Wang
%All Rights Reserved.
%
%The author was with Howard Hughes Medical Institute, and Laboratory
%for Computational Vision at Center for Neural Science and Courant
%Institute of Mathematical Sciences, New York University, USA. He is
%currently with Department of Electrical and Computer Engineering,
%University of Waterloo, Canada.
%
%----------------------------------------------------------------------
%Permission to use, copy, or modify this software and its
documentation
%for educational and research purposes only and without fee is hereby
%granted, provided that this copyright notice and the original
authors'
%names appear on all copies and supporting documentation. This program
%shall not be used, rewritten, or adapted as the basis of a commercial
%software or hardware product without first obtaining permission of
the
%authors. The authors make no representations about the suitability of
%this software for any purpose. It is provided "as is" without express
%or implied warranty.
%----------------------------------------------------------------------
%
%This is an implementation of the algorithm for calculating the
%Structural SIMilarity (SSIM) index between two images. Please refer
%to the following paper:
%
%Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image
%quality assessment: From error measurement to structural similarity"
%IEEE Transactios on Image Processing, vol. 13, no. 4, Apr. 2004.
%
%Kindly report any suggestions or corrections to zhouwang@ieee.org
%
%----------------------------------------------------------------------
%
%Input : (1) img1: the first image being compared
%        (2) img2: the second image being compared
%        (3) K: constants in the SSIM index formula (see the above
%            reference). defualt value: K = [0.01 0.03]
%        (4) window: local window for statistics (see the above
%            reference). default widnow is Gaussian given by
%            window = fspecial('gaussian', 11, 1.5);
%        (5) L: dynamic range of the images. default: L = 255
```

```
%
%Output: (1) mssim: the mean SSIM index value between 2 images.
%            If one of the images being compared is regarded as
%            perfect quality, then mssim can be considered as the
%            quality measure of the other image.
%            If img1 = img2, then mssim = 1.
%        (2) ssim_map: the SSIM index map of the test image. The map
%            has a smaller size than the input images. The actual
size:
%            size(img1) - size(window) + 1.
%
%Default Usage:
%   Given 2 test images img1 and img2, whose dynamic range is 0-255
%
%   [mssim ssim_map] = ssim_index(img1, img2);
%
%Advanced Usage:
%   User defined parameters. For example
%
%   K = [0.05 0.05];
%   window = ones(8);
%   L = 100;
%   [mssim ssim_map] = ssim_index(img1, img2, K, window, L);
%
%See the results:
%
%   mssim                        %Gives the mssim value
%   imshow(max(0, ssim_map).^4)  %Shows the SSIM index map
%
%=====================================================================
if (nargin < 2 | nargin > 5)
   mssim = -Inf;
   ssim_map = -Inf;
   return;
end

if (size(img1) ~= size(img2))
   mssim = -Inf;
   ssim_map = -Inf;
   return;
end

[M N] = size(img1);

if (nargin == 2)
   if ((M < 11) | (N < 11))
        mssim = -Inf;
        ssim_map = -Inf;
      return
   end
   window = fspecial('gaussian', 11, 1.5);  %
   K(1) = 0.01;                             % default settings
   K(2) = 0.03;                              %
   L = 255;                                   %
end

if (nargin == 3)
   if ((M < 11) | (N < 11))
        mssim = -Inf;
        ssim_map = -Inf;
      return
   end
   window = fspecial('gaussian', 11, 1.5);
   L = 255;
   if (length(K) == 2)
      if (K(1) < 0 | K(2) < 0)
              mssim = -Inf;
           ssim_map = -Inf;
           return;
      end
   else
        mssim = -Inf;
     ssim_map = -Inf;
        return;
   end
end

if (nargin == 4)
   [H W] = size(window);
   if ((H*W) < 4 | (H > M) | (W > N))
        mssim = -Inf;
        ssim_map = -Inf;
      return
   end
   L = 255;
   if (length(K) == 2)
      if (K(1) < 0 | K(2) < 0)
              mssim = -Inf;
           ssim_map = -Inf;
           return;
      end
   else
        mssim = -Inf;
     ssim_map = -Inf;
        return;
   end
end

if (nargin == 5)
   [H W] = size(window);
   if ((H*W) < 4 | (H > M) | (W > N))
        mssim = -Inf;
        ssim_map = -Inf;
      return
   end
   if (length(K) == 2)
      if (K(1) < 0 | K(2) < 0)
              mssim = -Inf;
           ssim_map = -Inf;
           return;
      end
   else
        mssim = -Inf;
      ssim_map = -Inf;
         return;
   end
end

C1 = (K(1)*L)^2;
C2 = (K(2)*L)^2;
window = window/sum(sum(window));
img1 = double(img1);
img2 = double(img2);

mu1   = filter2(window, img1, 'valid');
mu2   = filter2(window, img2, 'valid');
mu1_sq = mu1.*mu1;
mu2_sq = mu2.*mu2;
mu1_mu2 = mu1.*mu2;
sigma1_sq = filter2(window, img1.*img1, 'valid') - mu1_sq;
sigma2_sq = filter2(window, img2.*img2, 'valid') - mu2_sq;
sigma12 = filter2(window, img1.*img2, 'valid') - mu1_mu2;

if (C1 > 0 & C2 > 0)
   ssim_map = ((2*mu1_mu2 + C1).*(2*sigma12 + C2))./((mu1_sq + mu2_sq
+ C1).*(sigma1_sq + sigma2_sq + C2));
else
   numerator1 = 2*mu1_mu2 + C1;
   numerator2 = 2*sigma12 + C2;
      denominator1 = mu1_sq + mu2_sq + C1;
   denominator2 = sigma1_sq + sigma2_sq + C2;
   ssim_map = ones(size(mu1));
   index = (denominator1.*denominator2 > 0);
   ssim_map(index) =
(numerator1(index).*numerator2(index))./(denominator1(index).*denomina
tor2(index));
   index = (denominator1 ~= 0) & (denominator2 == 0);
   ssim_map(index) = numerator1(index)./denominator1(index);
end

mssim = mean2(ssim_map);

return
```

Appendix C

Journal Publications

Journal Publication 1: Anthony Winterlich, Ciaran Hughes, Liam Kilmartin, Martin Glavin, Edward Jones; “An oriented gradient based image quality metric for pedestrian detection performance evaluation.” *Signal Processing: Image Communication.* Vol. 31, February 2015, Pages 61-75

Journal Publication 2: Anthony Winterlich, Patrick Denny, Liam Kilmartin, Martin Glavin, Edward Jones; “Performance optimization for pedestrian detection on degraded video using natural scene statistics.”*SPIE Journal of Electronic Imaging.* Vol. 23 Issue 6, 2014

Journal Publication 3: Shane Tuohy, Anthony Winterlich [*], Brian McGinley, Martin Glavin, Edward Jones, Patrick Denny, Liam Kilmartin; “Evaluating the influence of packet loss on visual quality of perception for high bandwidth automotive networks.”*Signal Processing: Image Communication.* Vol. 43, April 2016, Pages 15-27

[*] Denotes corresponding author